THE UNIVERSITY *of* EDINBURGH

Proteins, Anatomy and Networks of the Fruit Fly Brain

Seymour Knowles-Barley

Doctor of Philosophy

Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh

2012

Abstract

Our understanding of the complexity of the brain is limited by the data we can collect and analyze. Because of experimental limitations and a desire for greater detail, most investigations focus on just one aspect of the brain. For example, brain function can be studied at many levels of abstraction including, but not limited to, gene expression, protein interactions, anatomical regions, neuronal connectivity, synaptic plasticity, and the electrical activity of neurons. By focusing on each of these levels, neuroscience has built up a detailed picture of how the brain works, but each level is understood mostly in isolation from the others. It is likely that interaction between all these levels is just as important. Therefore, a key hypothesis is that functional units spanning multiple levels of biological organization exist in the brain. This project attempted to combine neuronal circuitry analysis with functional proteomics and anatomical regions of the brain to explore this hypothesis, and took an evolutionary view of the results obtained. During the process we had to solve a number of technical challenges as the tools to undertake this type of research did not exist. Two informatics challenges for this research were to develop ways to analyze neurobiological data, such as brain protein expression patterns, to extract useful information, and how to share and present this data in a way that is fast and easy for anyone to access.

This project contributes towards a more wholistic understanding of the fruit fly brain in three ways. Firstly, a screen was conducted to record the expression of proteins in the brain of the fruit fly, *Drosophila melanogaster*. Protein expression patterns in the fruit fly brain were recorded from 535 protein trap lines using confocal microscopy. A total of 884 3D images were annotated and made available on an easy to use website database, BrainTrap, available at fruitfly.inf.ed.ac.uk/braintrap. The website allows 3D images of the protein expression to be viewed interactively in the web browser, and an ontology-based search tool allows users to search for protein expression patterns in specific areas of interest. Different expression patterns mapped to a common template can be viewed simultaneously in multiple colours. This data bridges the gap between anatomical and biomolecular levels of understanding.

Secondly, protein trap expression patterns were used to investigate the properties of the fruit fly brain. Thousands of protein-protein interactions have been recorded by methods such as yeast two-hybrid, however many of these protein pairs do not express in the same regions of the fruit fly brain. Using 535 protein expression patterns it was possible to rule out 149 protein-protein interactions. Also, protein expression patterns registered against a common template brain were used to produce new anatomical

breakdowns of the fruit fly brain. Clustering techniques were able to naturally segment brain regions based only on the protein expression data. This is just one example of how, by combining proteomics with anatomy, we were able to learn more about both levels of understanding.

Results are analysed further in combination with networks such as genetic homology networks, and connectivity networks. We show how the wealth of biological and neuroscience data now available in public databases can be combined with the BrainTrap data to reveal similarities between areas of the fruit fly and mammalian brain. The BrainTrap data also informs us on the process of evolution and we show that genes found in fruit fly, yeast and mouse are more likely to be generally expressed throughout the brain, whereas genes found only in fruit fly and mouse, but not yeast, are more likely to have a specific expression pattern in the fruit fly brain. Thus, by combining data from multiple sources we can gain further insight into the complexity of the brain. Neural connectivity data is also analyzed and a new technique for enhanced motifs is developed for the combined analysis of connectivity data with other information such as neuron type data and potentially protein expression data.

Thirdly, I investigated techniques for imaging the protein trap lines at higher resolution using electron microscopy (EM) and developed new informatics techniques for the automated analysis of neural connectivity data collected from serial section transmission electron microscopy (ssTEM). Measurement of the connectivity between neurons requires high resolution imaging techniques, such as electron microscopy, and images produced by this method are currently annotated manually to produce very detailed maps of cell morphology and connectivity. This is an extremely time consuming process and the volume of tissue and number of neurons that can be reconstructed is severely limited by the annotation step. I developed a set of computer vision algorithms to improve the alignment between consecutive images, and to perform partial annotation automatically by detecting membrane, synapses and mitochondria present in the images. Accuracy of the automatic annotation was evaluated on a small dataset and 96% of membrane could be identified at the cost of 13% false positives.

This research demonstrates that informatics technology can help us to automatically analyze biological images and bring together genetic, anatomical, and connectivity data in a meaningful way. This combination of multiple data sources reveals more detail about each individual level of understanding, and gives us a more wholistic view of the fruit fly brain.

Acknowledgements

Thanks to Douglas for always giving good advice and for steering me in the right direction. Thanks to Sarah, Bev, Ray and Flynn for their support and encouragement. Thanks to all members of the Armstrong, Meinertzhagen and Jarman labs for showing me the ropes, and also to past and present members of the iANC and Neuroinformatics DTC for listening, giving advice and helping make this project such an enjoyable experience.

Many thanks to the funding bodies that have supported this research, primarily to the Engineering and Physical Sciences Research Council for their support throughout this research. Many thanks also for funding and travel support from the Medical Research Council, the British Society for Developmental Biology and the Society for Experimental Biology. Thanks also to the Edinburgh University iDEA lab for GPU computing resources and funding.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Seymour Knowles-Barley*)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Neuroscience has traditionally taken a reductionist approach to understanding the brain. For example, after many years of research, neurons are categorized into different types by their shape and patterns of activity, regions of the brain are identified by their function, and many psychiatric and neurological disorders can now be classified based on underlying biological conditions. Research is usually focused on understanding one aspect of brain function in detail. This approach has been effective despite limited experimental techniques available for measuring brain functions. Biological and genetic techniques have also allowed neuroscience to investigate the biological functions that govern brain processes and link individual genes to important brain functions including development, plasticity, and disease. Contemporary advances in genetic and biological techniques now produce large amounts of data and enable a broad, integrative approach to neuroscience research [65]. This wholistic approach requires neuroinformatics tools to bring together data from numerous sources in a meaningful way.

To understand the complexity of the brain, neuroscience uses various levels of detail for investigation. For example, in the brain of the common fruit fly, *Drosophila melanogaster*, we can investigate the level of molecular biology to understand how genes interact and make proteins to form brain cells. The anatomical level can be investigated with microscopy to observe brain regions, neural projections and connections between neurons. Observing at the level of behavior, researchers measure many fundamental functions of the brain such as learning and memory, or courtship. Between the biological processes and resulting behavior many intermediate levels exist, such as the anatomical areas of the brain, the connectivity between regions or individual nerve cells, and the electrochemical activity of neurons. Focused research of each of these modalities, from the function of genes to the behavior of an animal, has re-

vealed much about brain function, but in most cases knowledge is limited to each level in isolation. Each level of abstraction is informative in itself, but detailed knowledge of how all levels interact to form an entire brain and a behaving animal is beyond our current understanding. It is likely that a rich network of interactions between elements in all these levels of abstraction is important for brain function. This project investigates the hypothesis that functional units spanning multiple levels of biological organization exist in the brain.

In the last decades biological research has progressed rapidly. Since the sequencing of the fruit fly genome in 2000 [1], shortly followed by the draft of the human genome in 2001 [21], researchers have used the genome as a reference map on which to locate genes and pool research efforts around the world. Recently, genetic interaction and protein interaction networks have been added to this map to form a rich, interconnected understanding of how life works. The advances in biology have enabled the field of neurobiology to flourish by using the genetic map to identify genes relevant to the brain.

The two biological levels of genes and proteins are fundamentally linked by the genetic interactions which create proteins and regulate genes. The genetic sequence determines how proteins are constructed. Proteins interact with each other in specific ways, and in turn regulate the operation of genes to create further proteins. Similarly, the anatomical level of brain regions and connectivity is fundamentally linked by the physical location of the connected neurons. The research described in this thesis shows that with informatics technology it is possible to bring together biological and anatomical levels of understanding in a meaningful way, and how this combination gives us a more wholistic and detailed view of the fruit fly brain.

The levels of understanding depicted in Figure 1.3 are examined in this thesis. They are the levels of molecular biology (genes, proteins, and the interactions between them), anatomy (anatomical areas of the brain, gene and protein expression patterns, and protein localization), and connectivity (the branching network of projections and physical connections between neurons). I will show how we can use modern genetic techniques and microscopy combined with informatics technology to link information from each these levels into a network of information. Informatics tools help us to record and visualize these networks, and can be used to analyze the relationships present within them. There are several other levels at which we would like to understand the *Drosophila* brain which are also under active research. For example, electrophysiology (chemical and electrical signals) concerns primary methods of signal

propagation in the brain, and behavior (learning, memory, courtship etc) is the observable output of neural activity in the brain. While these levels are outside the scope of this report, they are also linked fundamentally to the levels of molecular biology, anatomy and connectivity, and can be combined with other layers to produce a fuller understanding of neurobiology.

In the following sections of this chapter I will briefly introduce aspects of the fruit fly brain and outline the main areas of research covered in this thesis.

1.1 Molecular Biology

Model organisms such as *Drosophila* are widely used for genetic research. Because of the common evolutionary origins of life, we can study genes and proteins in the fruit fly brain which are similar to genes and proteins in humans. Crucially, many of these genes and proteins carry out the same functions in the fly brain that they perform in the human brain. This allows researchers to study neurobiology in the fruit fly to further our understanding of the workings of the human brain and of brain disease. Broad investigations into genetic similarities have found as many as 77% of human disease genes are also in the fruit fly [32, 109], including genes involved in disorders such as Alzheimer's disease, Parkinson's disease and Huntington's disease [139]. Human synaptopathies, like Alzheimer's disease, Parkinson's disease and schizophrenia, are pathologies of synaptic function. In order to treat these diseases we first have to understand how the synapse works at the level of genes and proteins; and at this level the *Drosophila* has already proven to be a relevant model organism [101, 139, 26].

The powerful genetic toolkit available to *Drosophila* researchers, and fast life cycle between generations, allows us to perform genetic investigations quickly. Biology is fundamental to the workings of the brain, so genetic investigations are key to discovering mechanisms that lead to many neurological phenomena, such as development, and the synaptic plasticity that underlies learning. The similar genetics between fly and human for many neurological disorders [109, 139, 50] means we can use the fruit fly as a model organism to study the fundamental neurobiology of these diseases. Also, its small size means that with confocal microscopy we can record 3D images where every cell in the brain is visible. Progress is also being made in the field of electron microscopy (EM), where researchers aim to record the shape of every neuron, and the synaptic connections that the neurons make with each other.

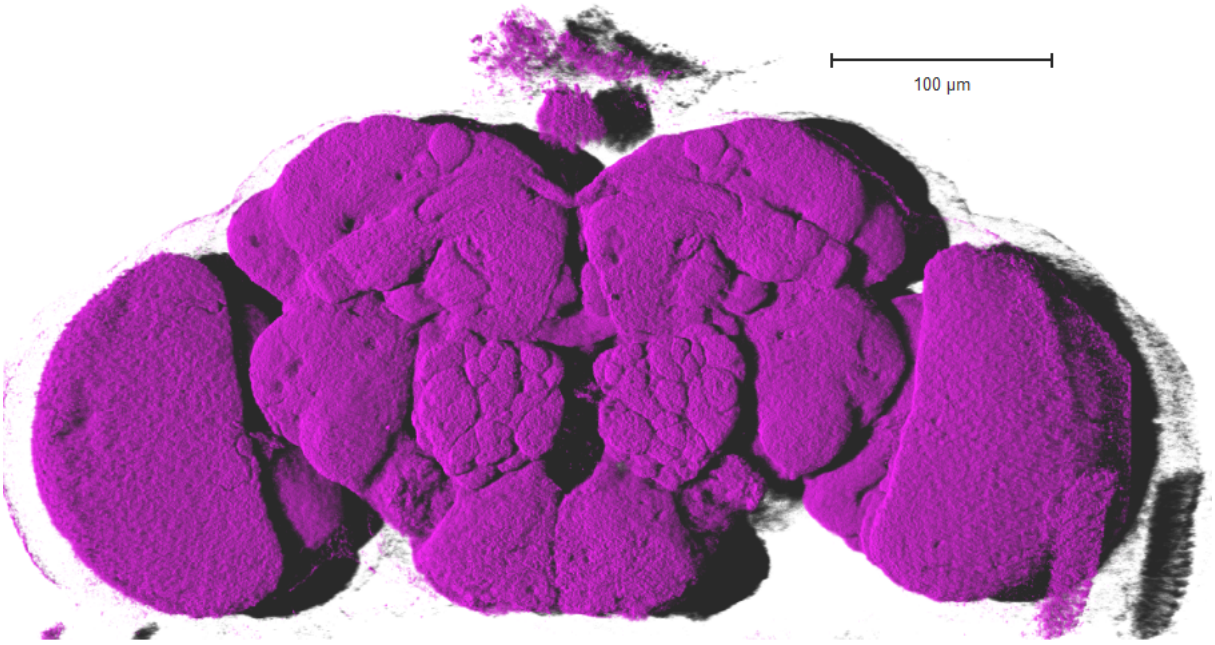


Figure 1.1: 3D fruit fly brain. This surface rendering of the *brp* expression in *Drosophila* brain is generated by combining images from a stack of 2D images. This view allows us to visualize the three-dimensional surface shape of the neuropil regions, but internal structures are hidden from view. The image is artificially coloured, with anti-*brp* signal in magenta.

1.2 Anatomy

Insect brain anatomy has been studied for many years. Some areas of the fruit fly brain have obvious functions, such as the optic lobes where input from the compound eyes is processed for vision, whereas other areas are less well defined and studied. Taking advantage of useful genetic techniques, functional studies can investigate the relationship between anatomy and behavior by locating brain areas required for specific actions [75, 86]. The stereotypical structure of the *Drosophila* brain means that gross anatomy is very similar for every fruit fly [18, 103]. This enables researchers to study the same neurons or anatomical regions in different fruit flies with genetic techniques.

The FlyBase anatomical ontology provides a common system of anatomy for the fly brain which was developed in collaboration with many groups in the fruit fly community [34, 5]. This common ontology is important for the research community as it allows databases to use consistent terms and will, in future, allow automatic queries to be generated to retrieve information from many different data sources. In Chapter 3,

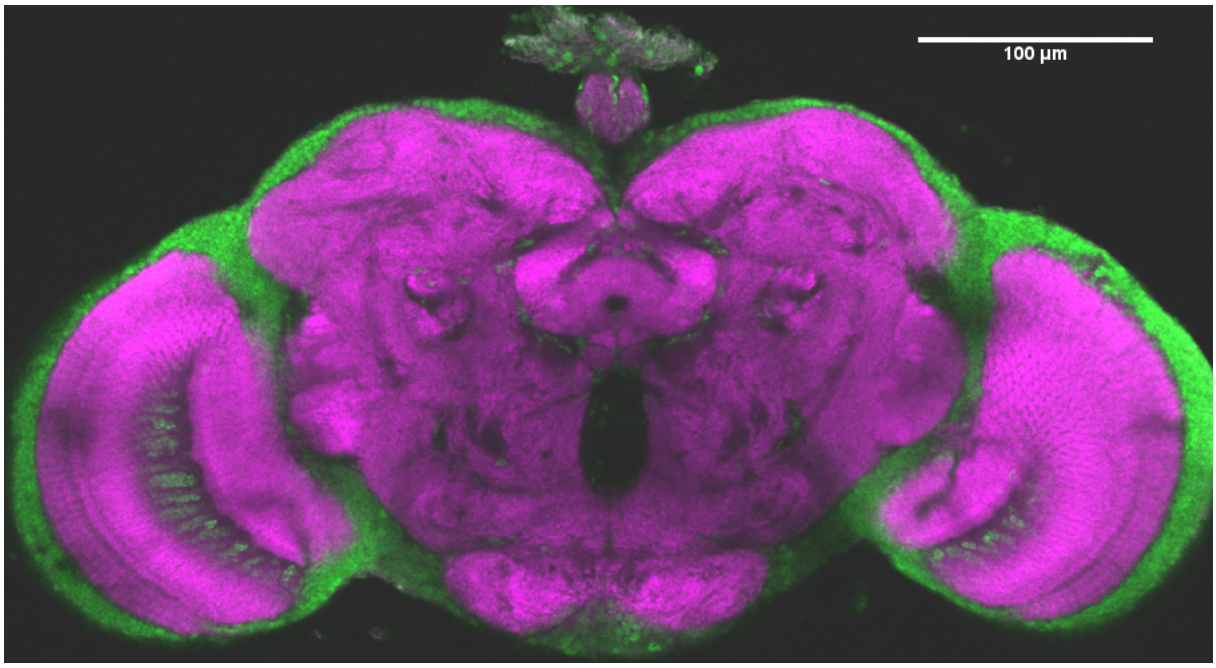


Figure 1.2: Single z image from the stack of images used to create Figure 1.1. This view allows us to identify internal structures within the brain. The image is artificially coloured to show anti-*brp* signal in magenta and EYFP expression in green. In this case the gene CG1910 is tagged with EYFP and the expression is seen only in the cell bodies, mainly forming a shell around the outside of the brain.

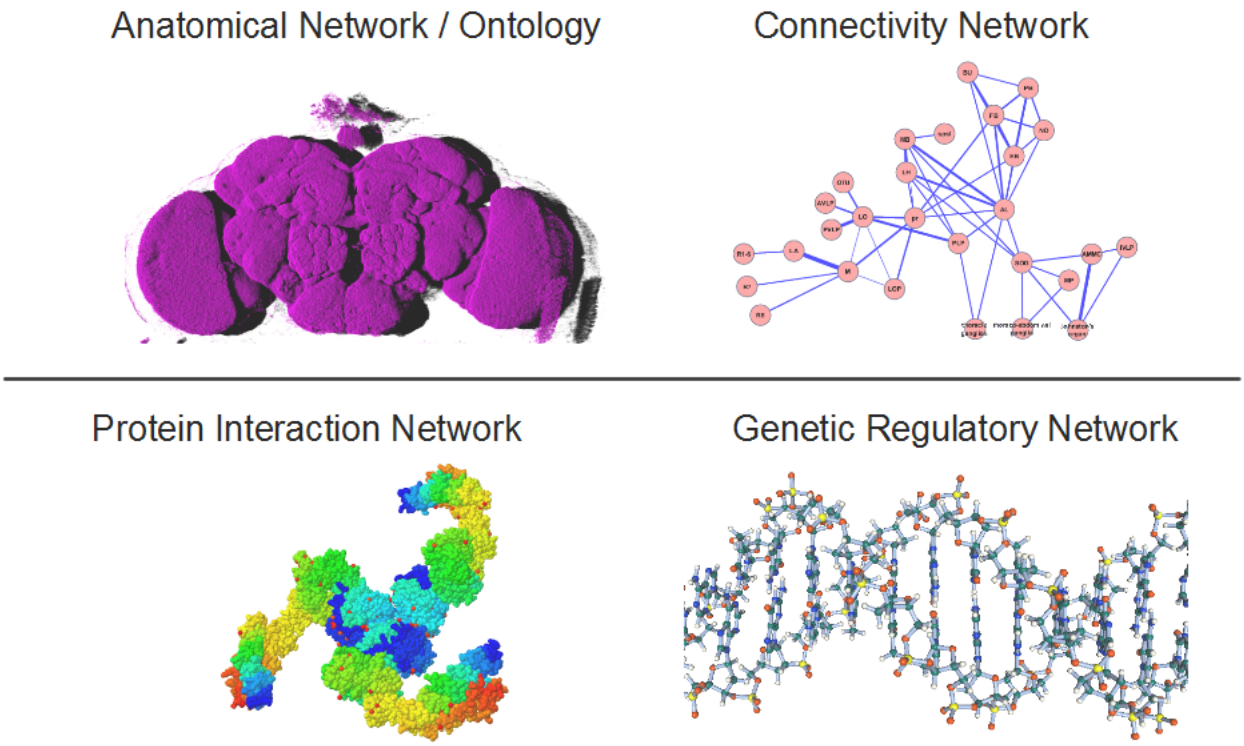


Figure 1.3: Brain networks studied in this project. Anatomy and connectivity networks are linked by the physical locations of neurons and connections. Protein and gene networks are fundamentally linked by the biological processes that create proteins and regulate genes. Interactions between all of these all these networks are likely to be important for brain function, but at present such interactions are not well understood.

protein expression patterns are imaged from protein trap fruit fly lines from the Cambridge Protein Trap (CPT) project, and annotated against brain areas from the FlyBase ontology. Such protein expression data gives us important information about how brain areas differ from each other, and provides a basis for further investigation and network analysis.

1.3 Networks

The way genes and proteins interact can be represented as a large interconnected network, where individual genes and proteins are linked together according to their interactions. Such networks allow us to describe and investigate the neurobiology of cells in the brain. Currently, website databases like FlyBase, Ensembl, NCBI etc describe

genes and proteins in great detail. Interactions between genes and proteins are also recorded in website databases such as the Drosophila Interaction Database (DroID); and gene homology databases, such as Ensembl Compara, link the same genes present in different species.

For neuroscience, we are also particularly interested in where and when these genes are expressed, and where the proteins are located in the brain. Unfortunately this information is currently lacking, and is usually discovered manually for genes and proteins of particular interest. This information is vital to collect; proteins are the fundamental building blocks of all cells. Protein localization patterns and interactions enact the biological "program" that, for example, wires up the brain, modifies synaptic connections between neurons to produce learning, and determines the excitability and firing properties of neurons.

In Chapter 3, I describe BrainTrap, the fruit fly brain protein database, which begins this collection of protein expression information in the fruit fly brain. Once represented in a database it is possible to further analyze this protein expression data. In Chapter 4, I show there are several ways to combine the BrainTrap data to identify spurious protein-protein interactions, naturally segment anatomical areas in the fruit fly brain, discover similarities between areas in fruit fly and mouse brains, and investigate the way that evolution has guided the variability of genes in the brain. These are just some of the ways we can now link information from various data sources, and apply network analysis techniques to gain a more integrative understanding of the brain.

1.4 Connectivity

An additional area which is important for understanding the brain is the connectivity of neurons in the brain, but unfortunately current research lacks any comprehensive databases of synaptic level neuronal connectivity, with the notable exception of *C. elegans* [136]. Due to the intricate branching structure of neurons in the brain, collecting and annotating this data has proved to be difficult to achieve.

To make a simple analogy, it is like attempting to record the three-dimensional structure of a large tangle of branches, grass and weeds (standing in for neuronal projections) in a particularly neglected part of a garden. In addition to recording the location and direction of every branch, twig and blade of grass, we want to also record where adjacent branches make contact (or neurons make synapses) with each other, and all this must be achieved without disturbing the arrangement of branches and con-

tacts. To complete this analogy for the fruit fly brain we would need to imagine the tangle of branches to contain around 100,000 wildly branching plants, each making contact with between 100 and 1,000 other branches, and reduce the size of the tangle so the branches have a diameter of around 100-20nm.

One way to tackle this difficult problem would be to colour each branch with a different dye, so they could be identified more easily, and indeed this is one method used in the fruit fly brain. Using genetic techniques it is possible to highlight small subsets of neurons, or individual neurons in the fruit fly brain [49, 140, 119]. And recently, further genetic techniques have allowed imaging of several neurons in different colours [36, 37]. Light based microscopy, such as confocal microscopy, is then used to image the brain and record the highlighted neurons. By repeating this procedure many times, it is possible to map individual images of neurons to a template and infer connectivity patterns for entire fruit fly brain. Unfortunately, fundamental limitations of light microscopy mean only a subset of neurons can be reconstructed in this way for each fruit fly brain [121, 19]. Also, synaptic connections between neurons are not visible in these images and can only be inferred or investigated using additional genetic techniques.

Returning to the branch analogy, another possible solution would be to immerse the complicated tangle in a fluid that will later solidify, such as concrete, or a thick cookie dough. After baking, the resulting solid block can be cut into thin sections and all aspects of the 3D structure can be observed. This approach is analogous to the use of electron microscopy, where brain tissues are embedded into a medium such as epoxy resin, and imaged at a resolution of up to 3nm per pixel. This approach has the possibility of measuring all connections between neurons in the reconstructed volume. Initial research into these techniques was done in the 1970s and 1980s when, using photographic plates, it was possible for small volumes to be reconstructed with a considerable amount of manual effort [136, 79, 84]. Technology now assists us in recording and storing images digitally and performing some reconstruction operations automatically, but major parts of the reconstruction process are still manual. This presents a challenge to informatics and computer science technologies to ease this bottleneck, and in Chapter 5, I outline one approach developed for improved alignment and semi-automatic annotation of these images.

1.5 Background

Neuroinformatics is a term that was coined relatively recently, and defines a field that combines neuroscience research with informatics technologies. Well before the term existed, neuroscientists were using mathematical and computational techniques. For example in 1949 Hebb used mathematics to postulate the now widely used Hebbian theory of synaptic plasticity [39], and in 1959 Hodgkin and Huxley calculated the conductance of ion channels in the squid giant axon [42]. At the time, computers were in their infancy and were not essential for the mathematical calculations involved.

Our recent ability to gather information in very large databases, such as results from genetic expression and interaction experiments, means that bioinformatics is now an essential part of genetic and biological research. Researchers can now share data throughout the world, and use computational techniques to make sense of the information they gather. Neuroscience research is also starting to take advantage of computational techniques to share data and help understand the brain. The project described in this thesis relies on informatics techniques to share and analyze data collected from the fruit fly brain. The BrainTrap database, for example, links to and integrates with other biological and neurological data sources from around the world.

Data collection methods, analysis techniques and computer algorithms developed in this thesis share some aspects or goals with other published work. In this section I will outline existing data resources, network analysis techniques, and image analysis methods that share components with the methods I have used for collecting and analyzing fruit fly brain data. Some examples use similar biological techniques, such as the protein trap method, to image protein localization, and some use similar computational techniques on different data.

1.5.1 Proteomics

Protein expression locations in the *Drosophila* brain are currently only known for a small subset of the *Drosophila* proteome and localization information is usually collected for investigations into a single protein, in a specific anatomical area. Common techniques for determining protein localization include antibody staining, as in [101], and targeted insertion of fluorescent protein into the genome to produce *in vivo* fluorescence as in [81]. Both these papers use the techniques to study protein localization in the *Drosophila* brain. Recent large scale projects undertaken by several universities aim to catalog protein expression locations for as many proteins as possible with the

use of fluorescent protein insertion techniques [6, 34, 113]. Rather than focusing on single proteins and single areas, these projects aim to record expression information for many areas of the *Drosophila* anatomy, at multiple stages of development.

Localization information will help to determine what biological functions each protein performs, and when combined into one large data set it will help determine which pairs of proteins can potentially interact with each other by providing co-localization information. The yeast two-hybrid (Y2H) method [29, 132, 102] has identified tens of thousands of potential protein-protein interactions listed online in databases such as DroID [142]. On further investigation many of these interactions may not be relevant for the brain, or may not occur at all *in vivo* [113, 107]. Protein expression data is one useful way to identify spurious interactions among the overly-connected interaction networks. Proteins that co-locate in the same areas *in vivo* can potentially interact to form protein complexes and proteins that do not co-locate are not able to interact with each other, so many possible protein-protein interactions can be ruled out using this co-location data, as demonstrated in Chapter 4, where 149 interactions can be ruled out between a subset of 535 proteins.

The confocal imaging I have undertaken is part of a large scale collaborative research project to collect expression location information for as many proteins as possible from protein trap lines generated by the CPT project as described in [113]. Large scale projects like the CPT project are much easier to carry out in small model organisms such as *Drosophila*. Brain research in *Drosophila* can also have important implications for humans because many neuron specific genes found in mammals have at least one homologue gene in *Drosophila* and gene functions are well conserved between a wide range of species, including humans. Olfactory learning and memory is a particular focus for this type of research because several olfactory learning genes are found to be common to both mammals and *Drosophila*, and the behavioral effects of these genes can be measured experimentally [24, 138].

Research into genetic diseases using *Drosophila* also has good potential for useful research because out of all genes known to affect human disease as many as 77% have orthologues in *Drosophila* [109]. Many other complex traits of human behavior have also been linked to neurological genes well conserved in *Drosophila*. Replacing a *Drosophila* gene with a human orthologue of the same gene has already proved useful in studying genetic influences on alcohol dependence, sleep disorders, Huntington's disease, and neurodegenerative diseases such as Alzheimer's and Parkinson's disease [80]. By forging links between the proteomics and anatomy of the *Drosophila* brain,

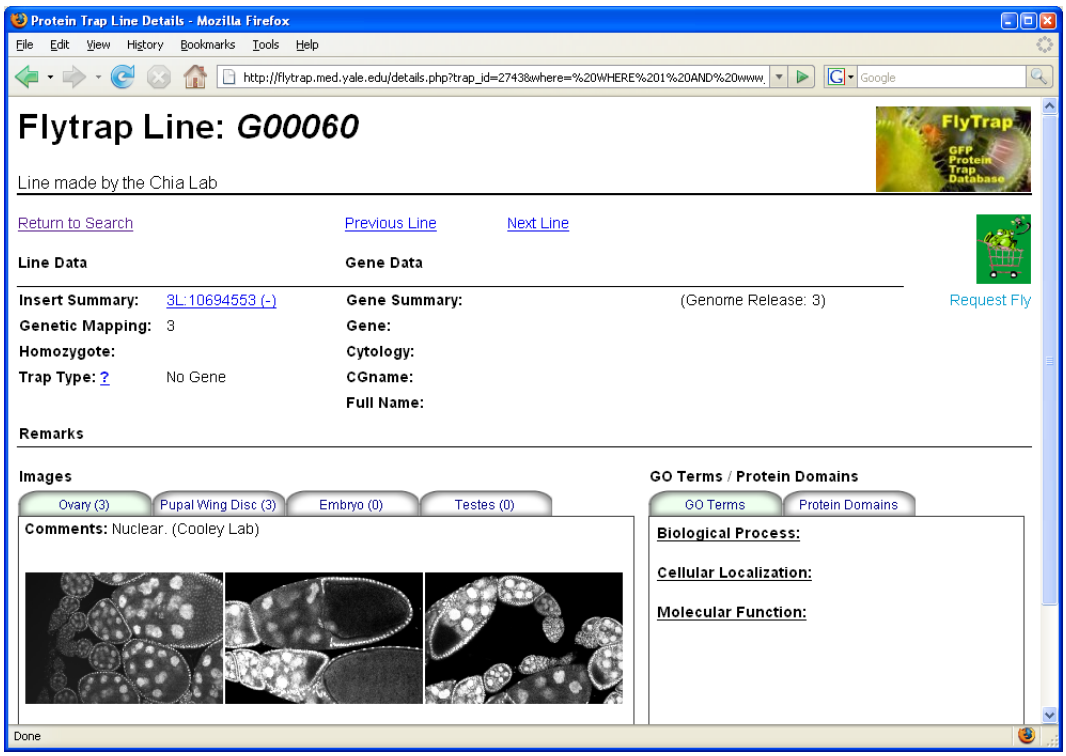


Figure 1.4: FlyTrap *Drosophila* protein expression database. In this view 2D images of protein expression in ovary and pupal wing discs can be displayed, for the selected protein trap line.

the BrainTrap database can provide information about the operation of proteins important for such disease research and for understanding the fundamental properties of the brain.

1.5.2 Online Databases

Collecting biological and anatomical data and sharing it over the internet is essential for both biological and neuroscience research. Several other databases and websites exist for collecting this type of information. Here I briefly describe some of these systems and compare the functionality they provide with the BrainTrap system for collecting fruit fly brain protein expression patterns.

1.5.2.1 FlyTrap

The FlyTrap database, http://flytrap.med.yale.edu, includes information from several large protein trap screens in *Drosophila*, and records protein localization patterns for several anatomical areas [59], not including the brain. FlyTrap also lists a wide range

Figure 1.5: FlAnnotator, the annotation website for the Cambridge Protein Trap (CPT) project. Annotated 2D images from many anatomical regions of protein trap fruit fly lines are listed in this database.

of information from the gene ontology consortium including information on biological processes, cellular components, and molecular functions of each protein. Free text searching is possible for all these fields, allowing researchers to search for protein trap lines of interest. 554 protein trap lines are included in the database, however currently only 184 lines have images associated with them. 2D images of protein expression are available online, from confocal images of ovaries, and pupal wing disks. Images from embryo and testes may also be included in the future.

FlyTrap shares a model organism and biological technique with the BrainTrap website, however the focus of the information being collected is different. The FlyTrap project aims to collect a very wide range of information about all protein trap lines so that any *Drosophila* researcher can find protein trap lines of interest. However, the image data available is currently limited to ovaries and pupal wing discs, whereas the CPT project (section 1.5.2.2) has images from a wider variety of tissues, and BrainTrap provides more advanced 3D image viewing functionality. 2D images are informative for the anatomical areas of interest in the FlyTrap database (ovaries, wing discs, embryo and testes) but because of the complicated 3D structure of the *Drosophila* brain it is not possible to display the full expression information in a single image.

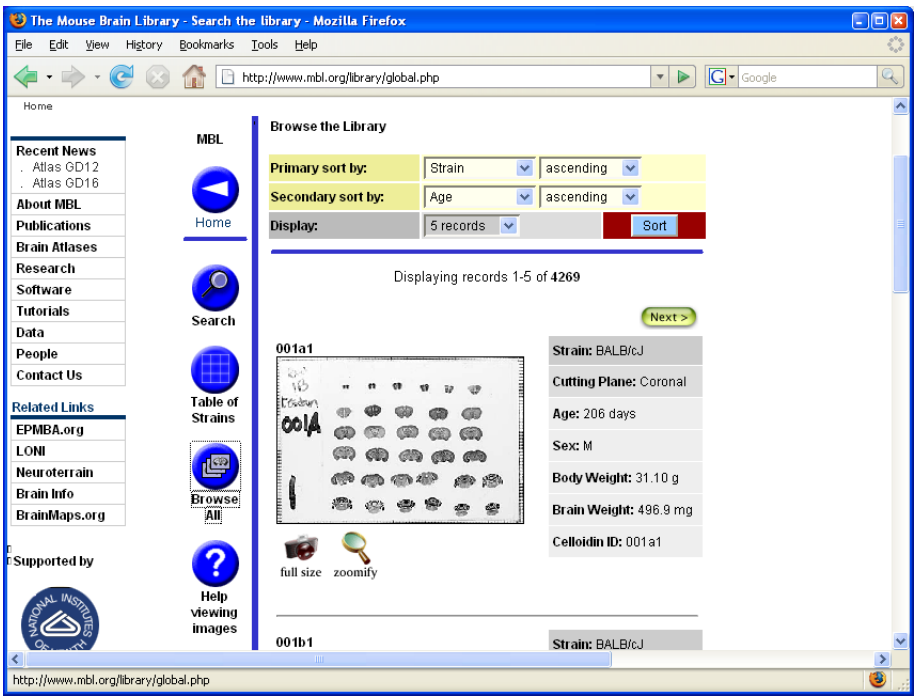


Figure 1.6: MBL annotated mouse brain database. In this search results view, multiple mouse brain slices can be viewed in a single 2D image.

1.5.2.2 Cambridge Protein Trap Project

A consortium of Universities from around the UK contribute to the Cambridge Protein Trap (CPT) project by imaging different anatomical regions of protein trap *Drosophila* lines generated at Cambridge University [112, 113, 107]. Image results are uploaded and made available on the FlyProt website, http://www.flyprot.org/. The BrainTrap website images are also available on the FlyProt site, where 2D images and expression summaries are shown. Again it is not possible to show the full 3D structure of brain expression patterns with single images, so the BrainTrap database provides 3D image viewing functionality for this data [63]. The FlyProt site contains details for each line including the insertion constructs used to generate each line, and sequencing information showing the exact location of the genetic insertions. Links from the BrainTrap website connect to the FlyProt site for this information.

1.5.2.3 Mouse Brain Library

The Mouse Brain Library (MBL), available online at http://www.mbl.org/, collects high resolution images of brains from many genetically characterized strains of mice [111]. The model organism and genetic techniques used by MBL differ from those

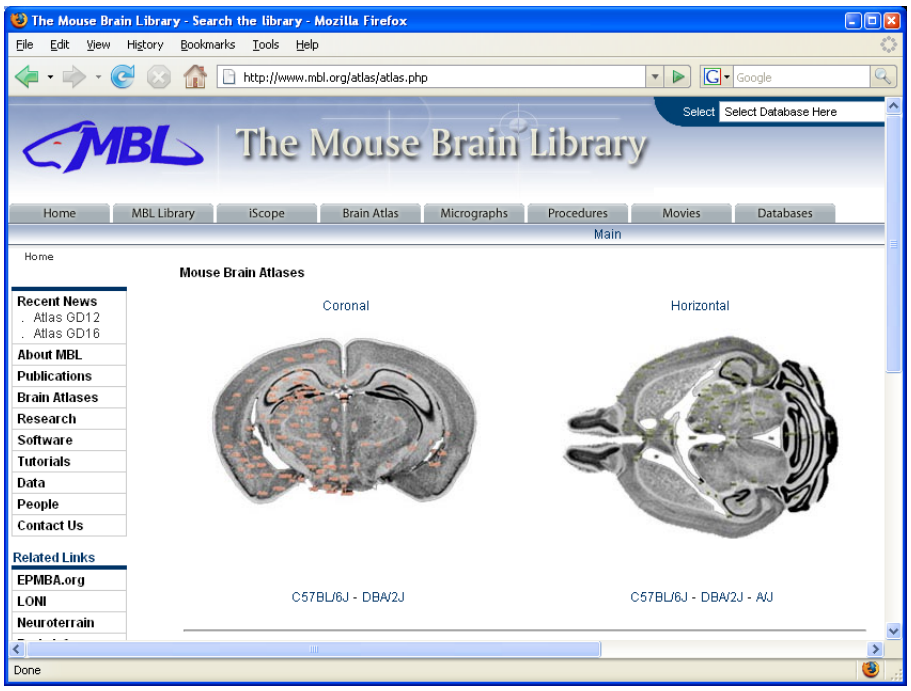


Figure 1.7: MBL annotated brain atlas. Stacks between 6 and 14 images have been annotated manually to indicate brain regions. Here the user can view annotated stacks in coronal of horizontal sections.

used in this thesis, however the informatics technology used to share the images over the internet serves the same purpose, and the MBL data also creates links between genetic and anatomical levels of understanding.

MBL does not store gene or protein expression information but does store images from many different species of mouse. Most strains of mouse are not aligned in a 3D stack, but rather displayed in one large 2D image (Figure 1.6). These images can be searched according to many physical characteristics of mice such as age, sex, weight etc.

Another section of the MBL website includes a small number of brain atlases, available at http://www.mbl.org/atlas/atlas.php. These are a set of images aligned in a stack that have been annotated to show the different brain regions in a series of 2D images (Figure 1.7). These stacks are easy to navigate, however each image stack only consists of between 6 and 14 images, whereas confocal images taken from the *Drosophila* brain can have between 100 to 200 images in a stack. A related site, the Electronic Prenatal Mouse Brain Atlas (EPMBA), available at http://www.epmba.org/, has many more images per brain, but navigation options are limited and it can be difficult to browse through the 3D stack in one page.

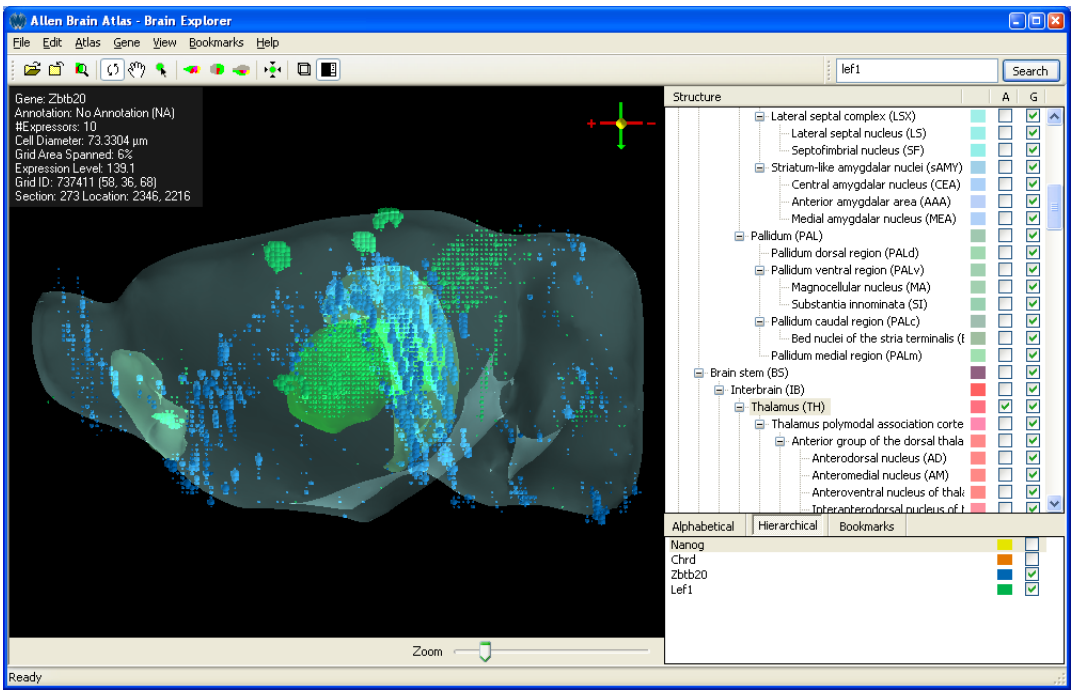


Figure 1.8: Allen Brain Atlas mouse brain 3D gene expression database. On the left, the brain viewing window allows the user to view 3D expression patterns interactively. On the right, an anatomical ontology allows the user to highlight brain regions of interest.

1.5.2.4 Allen Brain Atlas

The Allen Brain Atlas (ABA), available at http://www.brainatlas.org/, is a map of gene expression of over 20,000 genes in the mouse brain. Genes are mapped onto locations in a 3D template mouse brain, annotated and colour coded against an anatomical ontology [70].

The ABA requires a custom application (free to download) to view the template brain and the gene expression patterns in a 3D environment. Some of the database is also web-accessible, however for full functionality the application is easy to use and enables the brain to be viewed from any angle. The interface allows users to search for specific gene expression patterns and selected genes can be loaded into the viewer and displayed several at a time. The ABA website also allows searches for genes to be performed and the search form makes suggestions of possible gene names as you type. Once the expression data is loaded into the application, clicking on an area of expression retrieves the original 2D data that was used to generate any part of the 3D map (Figure 1.8). Because all expression patterns are mapped to a template brain it is easy to view several expression patterns at once and compare the expression locations. The template brain has anatomical regions classified already, so this allows automatic

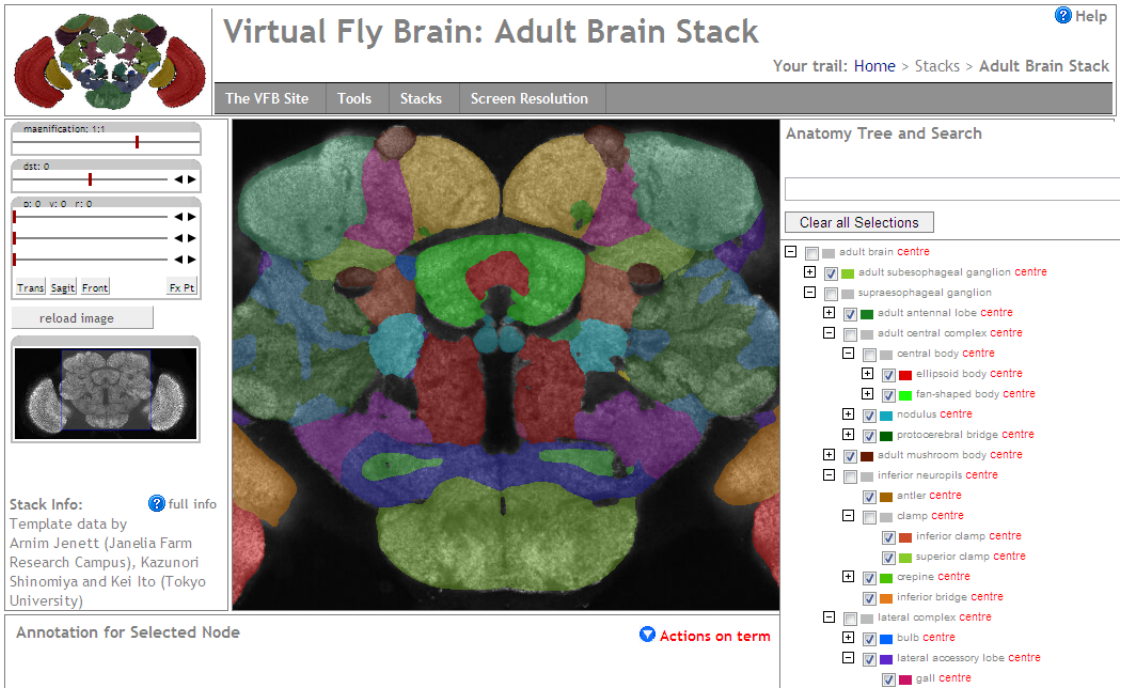


Figure 1.9: Virtual Fly Brain (VFB) website. On the left, view controls allow the user to change the 3D viewpoint and zoom level displayed in the viewing panel. In the center, the viewing panel shows the VFB template brain, with regions of the brain highlighted in different colours. On the right, the anatomical ontology allows the user to search for regions of interest, and to highlight them in the viewing window.

classification of brain regions once the gene expression data is mapped to the template. This is a very useful approach and is replicated in the BrainTrap website for a limited number of brain scans.

The ABA ontology tree interface is simple and easy to use, but it can be difficult to locate specific areas within the tree. BrainTrap tries to improve on this interface in a web based environment with the help of ontology term completion.

Compared to BrainTrap, the model organism and genetic techniques used by the ABA are different and the data distribution method requires a custom application. However, the goal of linking genetic data to anatomical locations is the same, and the anatomical resolution achieved is comparable between the two databases. In Section 4.3.1 these similarities are used to find correlations between BrainTrap protein expression and ABA gene expression data.

1.5.2.5 Virtual Fly Brain

The Virtual Fly Brain (VFB) website provides a 3D view of an annotated *Drosophila* brain and allows users to select regions of interest in the ontology to the right of the screen (see Figure 1.9). Regions are highlighted in the viewing window in the center of the screen. The brain ontology and region information was developed in collaboration with many *Drosophila* researchers from around the world including Edinburgh University, Cambridge University, Janelia Farm Research Campus, and Tokyo University. The VFB website demonstrates the anatomical ontology for the *Drosophila* brain and allows users to view many annotated regions. The website also allows the viewing plane to be adjusted to any angle. The anatomical ontology is the same as that used in BrainTrap, and is essential for consistent communication in the *Drosophila* research community. In the future, use of a common template throughout the research community will allow researchers to pool results quickly, and link multiple levels of understanding through resources such as the VFB website.

1.5.3 Network Analysis

Networks are a powerful tool for measuring complicated relationships and linking data across modalities, such as the genetic, proteomic, anatomical and connectivity levels considered in this thesis. Here I will discuss analysis methods already used for similar tasks, which are further developed in Chapter 4.

Data from the ABA has previously been used to investigate properties of different brain regions, to provide a more informed and detailed anatomical segmentation of the mouse brain [9], thereby linking genetic and anatomical levels in the mouse brain. Because protein trap expression data achieves a similar level of detail (number of voxels for a single brain scan) in fruit fly, as is available for mouse in the ABA data, methods used to analyze mouse gene expression patterns can be adapted to analyze fly protein expression patterns. In Chapter 4, I adapt the approach taken in [9] to segment regions of the brain based on fly protein expression patterns. In [9] k-means is used to cluster regions, and in Chapter 4, I demonstrate that Gaussian mixture models are more effective than k-means at segmenting fruit fly brain regions based on the protein expression data.

ABA data was also used to identify correlated genes with similar expression pattern profiles [90], to investigate the evolution of brain gene variability [26]. BrainTrap data can be analyzed in the same way, as shown in Section 4.3.

We can also use network analysis to combine data from ABA and BrainTrap databases. Online homology databases such as Ensembl Compara and NCBI homologene, available at:

- http://www.ensembl.org/info/docs/compara/
- http://www.ncbi.nlm.nih.gov/homologene

Respectively, offer a way to connect the data in the BrainTrap website with data in the ABA. Using protein expression patterns in BrainTrap and their homologus gene expression patterns in the ABA to discover correlations between fly brain and mouse brain regions. And using methods similar to [26] section 4.3.2 investigates the evolution of brain proteins using the BrainTrap protein expression patterns.

Connectivity data has also been analyzed in many ways before, and motif analysis is a well established method for comparing networks of different types to discover similar fundamental building blocks [87]. Using connectivity data from the *C.elegans* connectome motif analysis has also been used to search for computational modules in neural systems [108]. In section 4.4.1 this method is extended to incorporate neuron type information for *Daphnia magna* and *Drosophila* lamina data. This is one way that connectivity data could be combined with the BrainTrap protein expression data. Connectivity data is currently not available for the full fruit fly brain, but this analysis may be possible in future.

1.5.4 Template Registration

The stereotypical anatomy of the fruit fly brain allows us to compare brain images from different fruit flies directly against each other. Natural biological variability and variability introduced by genetic techniques, dissection and staining of individual brains for scanning means that images must be processed to allow direct comparison. This is very important for brain research as it allows us to link results from multiple experiments to the same anatomical locations.

One method often used to automate confocal image analysis is to use a template brain image that has already been manually annotated [51]. Several software packages are available to warp fruit fly brain images to a template, including the Computational Morphometry Toolkit (CMTK) [110] and BrainAligner [103]. CMTK was used to create the template registered brains on the BrainTrap website. Both systems are outlined briefly in the following sections. The general approach is to match new brain images

to the template by stretching so that the borders and detailed 3D structure of the new image matches those in the template. Distortions are applied to the image to get a better fit to the template. This approach is successful for comparing two similarly shaped brains, and can be used in conjunction with an annotated template brain to determine brain areas automatically. Repeating this process several times for different brain images it is possible to compare expression patterns directly. I demonstrate the use of this approach on the BrainTrap website, where a selection from 40 protein expression patterns can be merged into one single image.

1.5.4.1 CMTK

This open source, freely available software package consists of a collection of command-line tools for 3D image registration [110]. It has been applied successfully to 3D CT scans of human brain and 3D confocal images of insect brains, including *Drosophila*.

CMTK works by maximizing the mutual information between a new brain image and a template brain image. In most cases the principal components of the image are used to determine an initial affine transformation, which is then improved to find an affine transformation that maximizes the mutual information between the transformed image and the template image. Then a 3D grid of warp points are evenly distributed throughout the image and grid point locations are adjusted to create a B-spline deformation that again maximizes the mutual information between warped image and template. The grid is initially coarse, with perhaps 20-100 grid points, and successive iterations use finer grids until the desired grid accuracy is reached.

The initial affine transform is important to get right because only minor corrections are possible during the warping step. In most cases the principal components are sufficient to produce a good transformation, however in some cases manual adjustment of the initial affine transform is necessary.

Previous work on Drosophila brains has involved identifying anatomical differences between male and female brains [12] and mapping subsets of neurons traced using fluorescent markers to the same template brain to investigate regional connectivity [49].

CMTK was used to perform brain registration on BrainTrap images in Chapter 3. Details of CMTK methods can be found in Section 2.4.2.

1.5.4.2 BrainAligner

The BrainAligner software [103] is available as a plug-in for V3D [105], which is a set of software tools for confocal image analysis and annotation.

BrainAligner works by comparing selected landmark points in the template image to points in the new brain image to be warped. Comparisons between landmark points and points in the new image are made using a set of metrics: mutual information, inverse intensity difference, correlation, and similarity of invariant image moments. A consensus is calculated from these metric scores to determine the matching point in the new image. A thin plate spline deformation is then calculated to transform the new image to the template.

The software has been used to register many neurons to a single template image to investigate regional connectivity in the *Drosophila* brain [103, 104].

The BrainAligner system requires many manually selected landmark points in the template brain. For example, 172 points were used in [103]. CMTK does not require any manually selected landmarks. For this reason I used CMTK for template registration in Chapter 3.

1.5.5 EM Reconstruction

Neuroscience knowledge is severely limited by the complicated experimental procedures involved in measuring neuronal activity and the connections between neurons. Connectivity is often inferred by measuring the correlated activity of neurons, but the true nature of the full connectivity pattern between cells is very difficult to measure.

To reconstruct neural circuits from serial section transmission electron microscopy (ssTEM) images it is first necessary to align consecutive 2D EM images. By repeating this process over a collection of 2D images, a 3D image can be generated. Then by identifying outlines of cell membranes in each of the 2D images, segments of neural projections are identified, and by joining these in consecutive images it is possible to generate a 3D reconstruction of neurons and to determine the pattern of neural connectivity. 2D images are often aligned manually, with the help of some alignment software, and cell boundary identification is usually performed manually by an expert user. Alignment and reconstruction of the EM images is very time consuming, even when using software to assist with the alignment. Tools such as *IMOD* [66] and *Reconstruct* [28] are commonly used by researchers to assist in the alignment and reconstruction process, but these tools are primarily used to record and database information entered

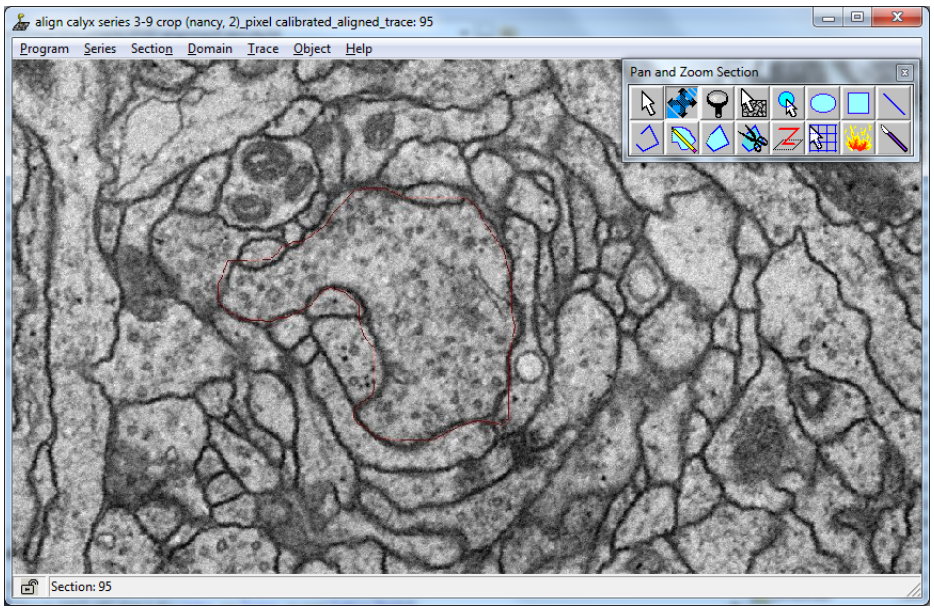


Figure 1.10: Reconstruct [28] screen shot with showing data from Ian Meinertzhagen's lab with annotation from Nancy Butcher [126, 62].

by the user, without further analysis or optimization [10]. Some attempts have recently been made to align stacks of EM images automatically and results suggest it is possible to improve upon manual alignment [23, 115], however small alignment errors still exist in most cases. In Chapter 5, I describe an automatic alignment system that can improve on existing alignments made either manually or by automatic landmark detection systems.

Computer vision algorithms have been also applied to EM data to aid researchers in annotation. However, depending on the image quality and methods of preparation and imaging, accuracy is limited and many hours of annotation are still required in most cases. Current algorithms used to perform this analysis repeat 2D operations many times to build up a 3D volume. Partly this is because EM data has a high resolution in two dimensions but a low resolution in the third dimension. In section 5.4 I show that by adapting 2D algorithms, full 3D analysis is possible for some high resolution data obtained using a focused ion beam / scanning electron microscope (FIB/SEM).

There are many software systems available for image alignment, annotation and segmentation of neurons in EM images. In the following sections I will discuss some commonly used packages.

1.5.5.1 Reconstruct

Reconstruct is a free, open source software tool for computer assisted manual annotation of EM volumes [28]. It is widely used in the community and can handle very large volumes of data. Images can be manually aligned using landmarks to identify alignment points in consecutive images, and any structure of interest can be manually annotated in 3D. The standard method of tracing a neuron is to manually draw around the outline of the neuron profile (Figure 1.10), for every image in the stack. After repeating this for all neurons or objects of interest a 3D view can be generated. Reconstruct is very flexible, and can be used to annotate any structures of interest, however the manual annotation is very time consuming. This software has been used to perform many EM analysis annotation projects including *Drosophila* neurite reconstructions [126, 71], mammalian spine and synapse reconstructions [56, 85, 22], and many more.

1.5.5.2 TrakEM2

TrakEM2 is a free, open source plugin for ImageJ [115, 15]. It performs alignment operations including microscope lens correction automatically. Many image processing tools built into ImageJ can also be used, for example a random forest classifier can perform trainable image segmentation [57] and there are tools for 3D visualization [117]. Neurons can be traced by drawing around region profiles in each image or by tracing out the skeleton of each neuron. For the skeleton tracing method only the middle of each neuron profile is marked rather that the full profile shape. This means the full volumetric shape of the neuron is not preserved, but the connectivity can still be established with this method, and tracing is much faster (although still very time consuming). TrakEM2 is a very useful tool which can perform many operations automatically, semi-automatically, or manually. The flexibility in tracing style and number of analysis methods available means that the interface can be difficult to learn and it is not obvious how to use the software initially. TrakEM2 has been used to generate several large scale connectivity reconstructions including reconstructions from mouse cortex [8] and *Drosophila* larva [15, 14]. A compatible tool for collaborative annotation using a web browser, named CATMAID, is also under development [114].

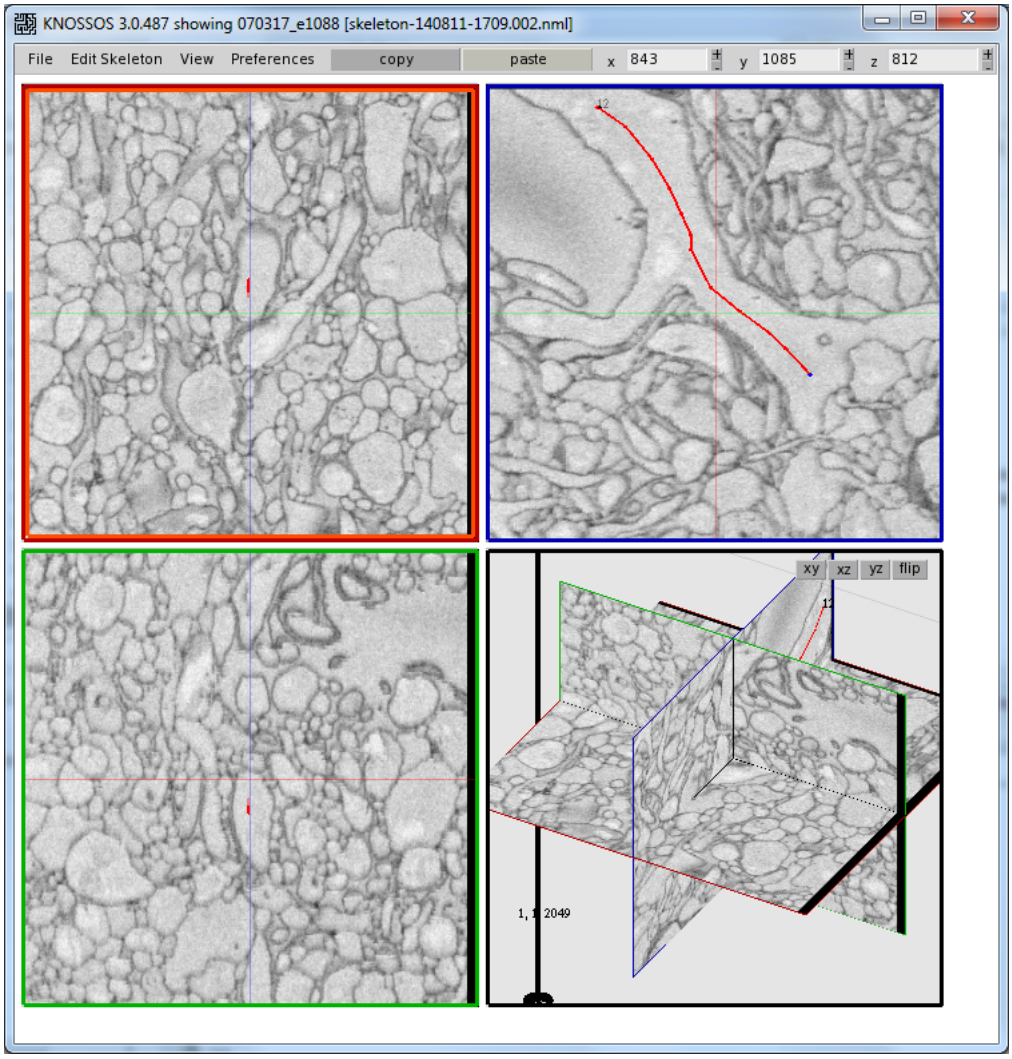


Figure 1.11: KNOSSOS screen shot showing example SBFSEM data from the KNOSSOS website [41]. The red line in the upper right panel shows a partially traced skeleton of a neurite.

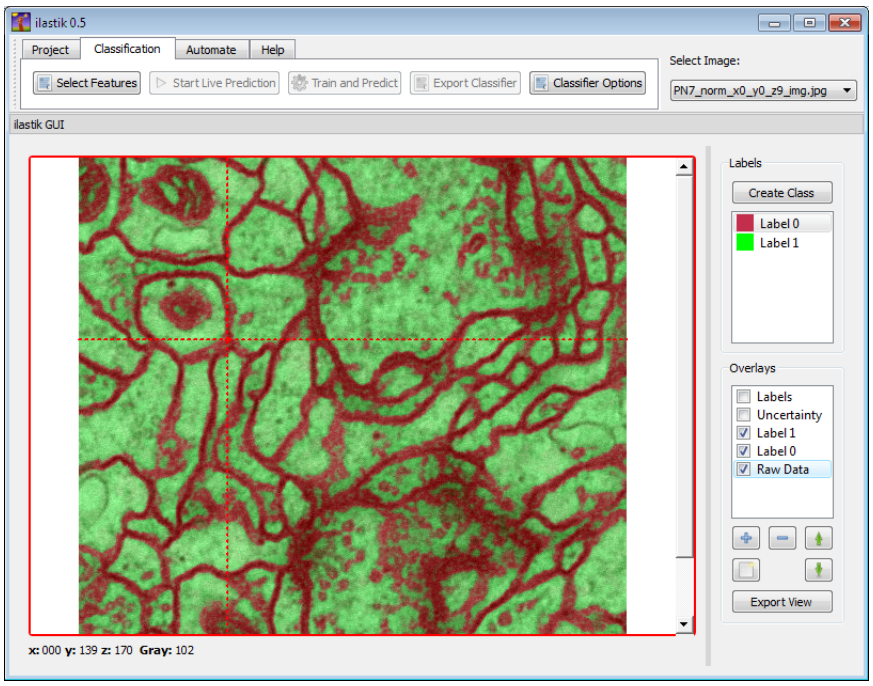


Figure 1.12: Ilastik screen shot showing segmentation results for an image from the data used in [62].

1.5.5.3 KNOSSOS

An annotation system named KNOSSOS [41] has been optimized for the skeleton tracing method and for serial block face serial electron microscope (SBFSEM) data. The three panel view shown in Figure 1.11 allows the user to view the neurite profile from three different angles at once. This view is more practical for isotropic, or near isotropic data (where xy resolution is similar to z resolution) such as the SBFSEM example data provided on the KNOSSOS website (Figure 1.11). Data obtained from methods such as ssTEM are anisotropic, with a z resolution approximately 15 times lower than the xy resolution, so tracing is usually performed in a single panel view. The KNOSSOS system also allows many users to perform annotations at once. Another software tool, redundant-skeleton consensus procedure (RESCOP), was developed to merge annotation results into consensus traces and identify areas of disagreement between different annotators [41]. KNOSSOS and RESCOP have been used to reconstruct large connectivity networks in the mammalian retina [11].

1.5.5.4 ilastik

The ilastik tool provides a targeted interface and very useful trainable segmentation system that uses a random forest classifier to label images based on user selected examples. In the example in Figure 1.12 the system has been trained to recognize membrane in EM images, however the system is very flexible and can be adapted to many automated labeling tasks [123] . The segmentation performance is very good, and was used to benchmark the segmentation system described in Chapter 5. By default ilastik uses generic features for training. When modified to use receptive fields generated in Chapter 5, EM image segmentation was improved [62].

1.6 Summary

The combination of proteomics data with anatomical specificity and connectivity information offers an opportunity to investigate hidden complexity in the neurobiology of the brain. Functional units spanning multiple levels of complexity are likely to exist between many levels in the brain, including the levels of molecular biology, anatomy and connectivity. To link these levels I measured protein expression patterns in the *Drosophila* brain, as described in Chapter 3, and combined results with multiple data sources available today to link the molecular, anatomical and connectivity, as described in Chapter 4. This analysis would not be possible without the aid of publicly available databases containing results from many thousands of biological and computational experiments. I also improve the automatic analysis of valuable EM image data to annotate neuronal connectivity information, as described in Chapter 5. With renewed interest in this area of connectomics, the development of more effective tools for this type of annotation is essential for future research.

Sharing protein expression data online, and integration of biological data from multiple sources required the development of new computational tools as, described in Chapter 3 and Chapter 4. This combination of computational analysis and experimental techniques is a powerful way to discover more about how the brain works.

Chapter 2

Materials and Methods

In this chapter I summarize the laboratory techniques, computational tools, and data sources used throughout the remaining chapters.

2.1 Protein Trap Lines

Fruit fly lines used in confocal and electron microscopy experiments in Chapter 3 were generated at Cambridge University as part of the CPT project [113]. The piggyBac insertion technique was used to randomly insert enhanced yellow fluorescent protein (EYFP), along with affinity purification epitopes FLAG and StrepII, into the *Drosophila* genome (Figure 3.1). The piggyBac method can insert sequences at random locations in the genome with different preferred insertion sites from other techniques [127, 43, 35].

2.2 Confocal Microscopy

Confocal microscope images are taken with a laser based microscope and can scan through small samples of tissue, such as a *Drosophila* brain, to generate a series of 2D images arranged into a 3D image stack. These images can be used to identify the location of fluorescent proteins and anatomical areas can be identified. Different structures within the *Drosophila* brain are visible using this approach and cell bodies can be identified. In some cases individual axon and dendrite tracts can be identified within the brain, but the fine structure of boutons and synapses is lost due to the resolution limitations of the confocal microscope.

In addition to the *in vivo* fluorescence of the endogenous EYFP, additional immunohistochemistry (IHC) was used to both enhance the EYFP signal, and to provide a background stain for brain structure identification and template mapping. *Drosophila* brains were prepared according to the following protocol:

Day 1 Dissect *Drosophila* brains under PBS (phosphate buffered saline).

Transfer dissected brains to 4% paraformaldehyde for one hour at room temperature.

Wash in PBS once for at least 5 minutes.

Wash in PBT (PBS with 0.3% by volume Triton X-100) three times for at least 20 minutes each time.

Incubate in primary antibodies overnight, for at least 20 hours (rabbit anti-GFP, Invitrogen, 1:1000 dilution, and mouse anti-*brp*, Developmental Studies Hybridoma Bank, 1:20 dilution).

Day 2 Wash in PBT three times for at least 20 minutes each time.

Incubate in secondary antibodies overnight, for at least 20 hours (goat anti-rabbit 488 and goat anti-mouse 546, both Invitrogen, 1:400 dilution).

Days 3-5 Wash in PBT for at least two days (change solution twice per day, at least 20 minutes apart).

Day 6 Wash in PBT one final time and mount brains in slides under 50:50 Vectashield and H_2O, using glycerol gelatin to secure cover slips to the slide.

Image using Zeiss LSM 5 Pascal confocal microscope.

The confocal microscope was set up in dual scanning mode, with laser emission at wavelengths of 488nm and 546nm, and two corresponding capture channels. When EYFP expression was particularly strong sequential scanning mode was used to eliminate any cross emission from the EYFP channel to the background channel. No cross emission was observed from the background channel to the EYFP in either scanning mode. In most cases a 20x objective lens was used and whole brains were scanned at a resolution of approximately 1024x512x120 voxels, with a voxel size of approximately 0.55x0.55x1µm.

Figure 1.2 shows some typical scans showing EYFP expression in green and background expression in magenta.

2.3 Electron Microscopy

Protein trap lines were also imaged using EM, in an attempt to identify the synaptic localization of trapped proteins. Several variations of preembedding (antibody staining is performed before embedding) and postembedding (antibody staining is performed after embedding) protocols were used, as described in Section 3.2. The main protocols followed for EM imaging are included here.

2.3.1 Preembedding

Drosophila brains were prepared for EM imaging using the following preembedding protocol.

Day 1 Dissect *Drosophila* brains under PBS (phosphate buffered saline) and fixative (well ventilated). Brains can be dissected, or alternatively the proboscis can be removed and the whole head can be fixed.

Transfer dissected brains to fixative (4% paraformaldehyde / 1% gluteraldehyde or 4% paraformaldehyde / 1% acrolein) in well plates for two hours.

Wash in PBT (PBS with 0.3% by volume Triton X-100) three times for at least 30 minutes each time.

Block with BSA.

Incubate in primary antibody for two nights, at least 40 hours (rabbit anti-GFP, Invitrogen, 1:1000 dilution) with NGS.

Day 3 Wash in PBT three times for at least 30 minutes each time.

Incubate in secondary antibody overnight, for at least 20 hours (goat anti-rabbit 1.5nm gold) with NGS.

Days 4 Wash in PBT at least three times for at least 30 minutes each time.

Silver enhancement reaction.

Fix in 1% osmium tetroxide (30 minutes).

Wash in H_2O (10 minutes)

Dehydrate in alcohol 50%, 70%, 90%, 100%, 10 minutes each.

Embed in epoxy resin, polymerize for 48 hours.

Day 6 Cut into 50nm sections using diamond knife ultramicrotome, collect slices on piloform grids.

Day 6-7 Image in electron microscope.

Results from preembedding experiments are shown in Figure 3.17. In this case, ultrastructure preservation was poor due insufficient fixation.

2.3.2 Postembedding

Drosophila brains were prepared for EM imaging using the following postembedding protocol.

Day 1 Dissect *Drosophila* brains under PBS (phosphate buffered saline) and fixative (well ventilated). Brains can be dissected, or alternatively the proboscis can be removed and the whole head can be fixed.

Transfer dissected brains to fixative (4% paraformaldehyde / 1% gluteraldehyde or 4% paraformaldehyde / 1% acrolein) for two hours.

Wash in PBT (PBS with 0.3% by volume Triton X-100) three times for at least 30 minutes each time.

Fix in 1% osmium tetroxide (30 minutes).

Wash in H_2O (10 minutes)

Dehydrate in alcohol 50%, 70%, 90%, 100%, 10 minutes each.

Embed in epoxy resin, polymerize for 48 hours.

Day 3 Cut into 50nm sections using diamond knife ultramicrotome, collect slices on piloform grids.

Day 4 Etch away resin from slices with acid (20 seconds).

Wash in H_2O (1 minute).

Block with BSA.

Incubate in primary antibody for 30 minutes (rabbit anti-GFP, Invitrogen, 1:1000 dilution with NGS).

Wash in H_2O (1 minute).

Incubate in secondary antibody for 30 minutes at room temperature (goat anti-rabbit 10nm gold with NGS).

Wash in H_2O (1 minute).

Stain with 2% uranyl acetate for 10 minutes at room temperature.

Image in electron microscope.

Results from postembedding experiments are shown in Figures 3.18 and 3.19. In this case, ultrastructure preservation was good, and some weak immuno-gold staining is visible.

2.4 BrainTrap Development

The BrainTrap website, http://fruitfly.inf.ed.ac.uk/braintrap/, was developed with the following software tools.

2.4.1 Website

The BrainTrap website was developed using Ruby on Rails server-side functionality, JavaScript to provide interactive browser elements, AJAX for dynamic communication between web pages and the server, XML for exporting database information. A PostgreSQL database was used to store all non-image information. Image files were extracted automatically from LSM files using Fiji at three different sizes for interactive viewing. Full size, 768 wide and 512 wide images were extracted. Each channel can be viewed separately or both viewed together in the interactive viewer.

The BrainTrap website uses the latest version of the FlyBase ontology [34] in all anatomy tags and on the search form.

2.4.2 MultiBrain

It was also possible to register some scans to a template image using existing software CMTK [110]. One confocal image of a male brain with a good nc82 channel signal

was selected from the BrainTrap data as a template image (Figure 1.2). Then a further 40 protein trap images with good nc82 channel signal, and distinct expression patterns were chosen and the nc82 channels were warped to the template brain. Using CMTK, an initial affine transform was generated based on image principal components to approximately match the template image by using the `make_initial_affine` command. The affine transform was then optimized to minimize the mutual information between the two images with the `registration` command. The CMTK `warp` command was then used to perform deformation adjustments to match the template exactly. In cases where warping was unsuccessful the affine transform was manually adjusted to better match the template image, and the `warp` command was repeated. Finally, protein signal channels were warped to the template using the same transformations.

The BrainTrap online viewer was adapted to allow simultaneous viewing of multiple protein trap channels that have been mapped to the template in this way. The viewer is available in the MultiBrain section of the BrainTrap website (Figure 3.14). Examples of several such multichannel images are shown in Figures 3.15 and 3.16.

In order to assess the accuracy of the template registration performed by CMTK, several pairs of registered protein expression patterns were compared to each other. Expression pattern pairs were chosen from different brain scans from the same protein trap line, or from expression patterns that highlighted the same anatomical structures. For each pair, easily identifiable points were highlighted manually in the non-warped version of the signal channel. For general expression patterns, annotations highlighted anatomically identifiable points such as the heel of the mushroom body or the most distal parts of the fan-shaped body. For sparse expression patterns annotations highlighted distinct features visible in both images such as cell bodies or bundles of neural projections.

Highlighted points were then warped to the template brain, using the previously computed warping functions based on the corresponding nc82 channels. Resulting warped annotations were compared directly to each other and the distance between highlighted points was measured. A total of 276 annotations were made on 14 brain images (7 pairs of matching expression patterns). The average distance between all highlighted points was $5.25\mu m$, with a standard deviation of $4.23\mu m$. This distance represents a combination of both natural and introduced variation between one brain image to another that cannot be corrected by CMTK. For example, we expect some natural variation in neuron locations to occur randomly during development, and some variation is unintentionally introduced by the dissection and scanning process.

Note that annotations made on cell bodies and parts of the adult brain cortex were significantly less accurate (average distance 7.34μm, s.d. 6.12μm) than annotations made inside the neuropil regions of the brain (average distance 4.49μm, s.d. 3.23μm, confidence level $\gamma = 0.9$). This suggests a higher variability in the location of cell bodies and brain cortex features compared to the location of neuropil regions and bundles of neural projections. However, an alternative explanation for this result is that the template registration is less accurate for cell bodies and adult brain cortex because they are located in a shell outside the nc82 expression pattern used to perform the alignment.

2.5 Density Development

The Density website described in section 5.3 was developed using Java and JPA server-side functionality, JavaScript and Raphael to provide interactive browser elements, AJAX for dynamic communication between web pages and exporting database information. A PostgreSQL database was used to store all non-image information including all trace annotations.

Density 2D tracing system was developed in Matlab and C++. Graphics processor unit (GPU) experiments were performed on an nVidia GTX 285, using Microsoft Visual Studio 2008 and nVidia CUDA SDK v3.2 and nVidia CUFFT libraries for fast Fourier transformations.

2.6 Data Sources

Data collected from external sources was used to analyze BrainTrap data in a wider context in Chapter 4, and as training and test data for neuron tracing algorithms in Chapter 5.

2.6.1 DroID

The Drosophila Interaction Database (DroID) [142] takes protein and gene interaction data from a wide variety of sources, such as Y2H, mass spectrometry, and homology data. A confidence score is calculated [141] to indicate the strength of evidence for that interaction. Interaction data form DroID (Data version 2011_05) was used to investigate protein-protein interactions with the BrainTrap data in Section 4.1.

2.6.2 Ensembl Compara

The Ensembl Compara website contains gene and protein homology data for many species, based on genome-wide species comparisons. Protein orthologue predictions between *Drosophila* and mouse from this database were used in Section 4.3 to find correlations between fly and mouse expression patterns. In Section 4.3.2, fly gene homologues in the yeast and mouse genomes were used to assess variability in ancient and recently evolved genes. The Ensembl Compara API was used to automatically download homology information for proteins in the BrainTrap database:

http://www.ensembl.org/info/docs/api/compara/index.html

2.6.3 Allen Brain Atlas

Three dimensional mouse brain gene expression patterns were used in Section 4.3.1 to compare fly and mouse brain expression patterns. The mouse expression data was obtained from the ABA [70], which is described briefly in section 1.5.2.4. Data was downloaded in XML format using the Allen Mouse Brain Atlas API:

http://mouse.brain-map.org/api/index.html

Where possible, expression patterns from multiple experiments were combined and the average expression level and density figures were used. In some cases ABA expression levels were found to be all zero or unavailable, and in other cases multiple expression profiles for the same gene were found to be very different. In such cases data for that gene was not used for further analysis.

2.6.4 EM Data

In Chapter 5, ssTEM images from a series of 50nm thick sections of the mushroom body calycal neuropil of *Drosophila* were used as training and test data for the EM segmentation and alignment improvement systems. The images were generated by Zhiyuan Lu and Ian Meinertzhagen, Dalhousie University, and manually annotated by Nancy Butcher, also of Dalhousie University see [71, 62] for details.

In Section 5.4.1 segmentation algorithms are further developed to utilize the extra information available in near isotropic EM images generated by the FIB/SEM method. Publicly available images provided by Graham Knott, EPFL, were used as training and test data [61, 78].

Chapter 3

Protein Trap Screen

Proteins are the building blocks of all cells and they play a key role in the structure and function of the brain. The location of particular proteins in the brain gives us important information about how areas of the brain differ and provides a useful set of data to help us understand how the brain is built, at the molecular level. This is particularly important at the synapse, where many proteins interact at the connection between neurons to form the post synaptic density. Research into synaptopathies, or diseases related to disfunction at the synapse, requires us to understand the functions of the synapse at the level of proteins. Synaptopathies already studied in the fruit fly include Alzheimer's disease, Huntington's disease, Parkinson's disease, and schizophrenia [80, 33].

To determine the location of proteins in the fruit fly brain researchers commonly use immunohistochemistry (IHC), which requires antibodies to be generated for the protein of interest. Because of the expensive and time consuming process of generating suitable antibodies these methods are usually only performed on proteins already identified as informative, so many potentially interesting proteins are ignored by these methods. An alternative approach is to modify the *Drosophila* genome by inserting a construct based on the green fluorescent protein (GFP) into an exon of an endogenous target protein. The intended outcome is that the target protein is expressed as usual, with a GFP protein attached, so that it will be visible under a fluorescence or confocal microscope. This is the approach taken in the CPT project [113, 112], where a piggyBac insertion method was used to randomly insert enhanced yellow fluorescent protein (EYFP), along with affinity purification epitopes FLAG and StrepII, into the *Drosophila* genome, as depicted in Figure 3.1. Affinity purification epitopes allow protein affinity capture and mass spectrometry to measure protein-protein interactions for the target protein [107]. Several other research groups have used also used variants

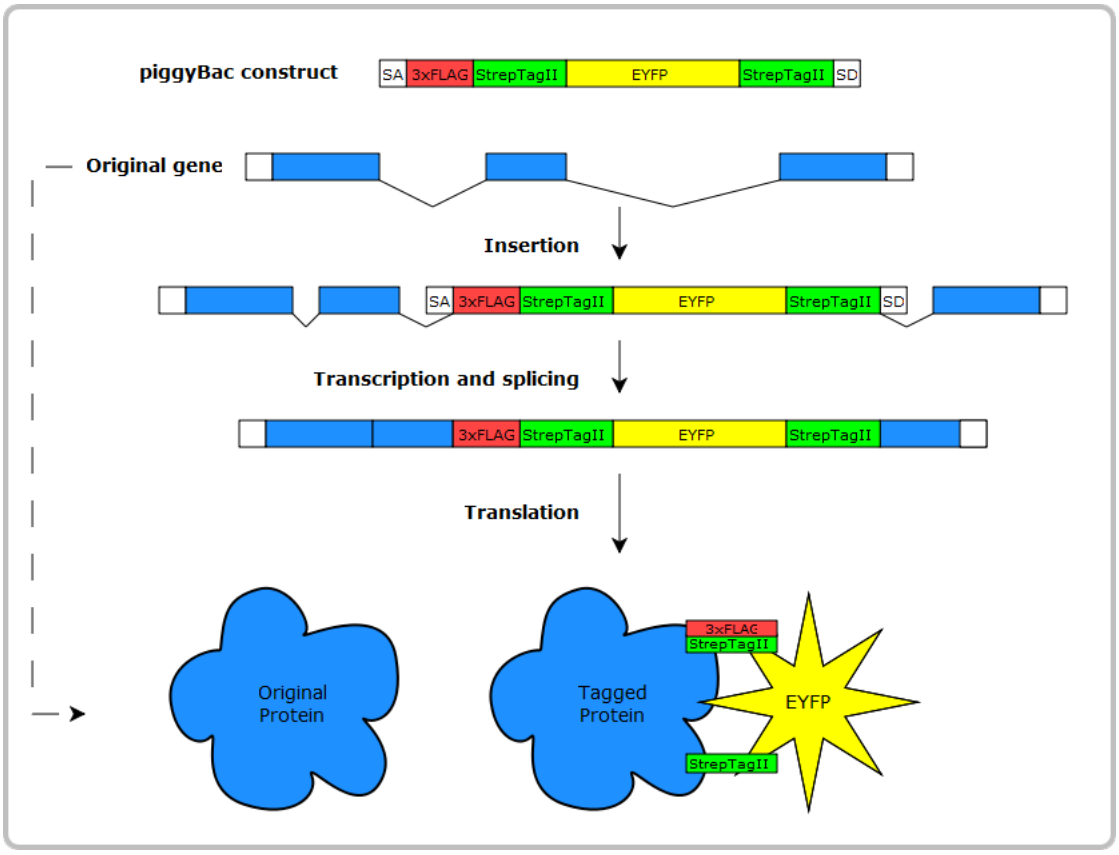


Figure 3.1: Schematic of the Cambridge Protein Trap (CPT) project insertion, using the piggyBac method. An artificial exon is inserted into the genome at random locations. When inserted within an intron of an existing gene the resulting tagged protein produces *in vivo* enhanced yellow fluorescent protein (EYFP) fluorescence and contains affinity purification epitopes StrepTagII and FLAG.

of GFP to trap as many proteins as possible, so that localization can be cataloged for a wider range of proteins, and so that protein interactions can be better understood. For example, the Yale FlyTrap trap project, described in section 1.5.2.1, includes images from *Drosophila* lines generated in labs around the world, including Yale University, the Carnegie Institution of Washington, and Kings College London.

3.1 BrainTrap

In order to assist in this research, and to provide an open, online information resource on protein expression in the fruit fly brain, I conducted a protein trap brain imaging screen using fruit fly lines from the CPT project. Brains from protein trap *Drosophila* lines generated by the CPT project were dissected and imaged using a confocal microscope. Images were annotated against the FlyBase anatomical ontology and an online database was developed to make the resulting images publicly accessible. The website, named BrainTrap, is available at http://fruit fly.inf.ed.ac.uk/braintrap, and currently contains protein expression information for 535 protein trap lines with 884 image stacks of protein expression patterns in the fruit fly brain. Full brain image stacks are viewable in a standard web browser and annotations mark expression in the brain against the FlyBase anatomical ontology. Annotation data can be used to search for proteins that express in a particular area of interest, or summarized to compare expression patterns of selected proteins side by side.

Some of this work also appears in the published paper "BrainTrap: a database of 3D protein expression patterns in the Drosophila brain", included as Appendix A.

3.1.1 Confocal Microscopy

Once the protein trap lines are generated, EYFP fluorescence can be observed directly under a fluorescence or confocal microscope to record the expression pattern of the trapped protein. However the level of brightness is dependent on the endogenous level of expression of the particular protein. Trapped proteins that are abundant in the brain can be imaged successfully without further staining. For other proteins the expression is low, making fluorescence levels weak and difficult to image clearly. Expression patterns alone can also be difficult to interpret without a background stain also present to give context and provide regional markers for the fine detail of different brain regions. For these reasons we used IHC to enhance the EYFP signal and to provide a

Protein	Lines	D.E.	Protein	Lines	D.E.
14-3-3epsilon	4	0	gish	1	0
Arm	1	0	nemy	1	0
Bsg	3	1	nrg	1	1
CaMKII	1	0	nrv2	1	0
cpx	1	0	orb2	1	0
dlg1	3	0	sgg	3	0
Fas2	2	1	tau	2	0
Fas3	3	1	zip	2	1
Gad1	1	0			
			Total	31	5

Table 3.1: Proteins trapped in the CPT project for proteins with expected expression patterns, as reported in existing literature. Column “Lines” lists how many protein trap lines have data available on the BrainTrap website and column “D.E.” shows the number of lines that show a different expression pattern than the one expected from previously reported results. Disagreement is observed in 16% of lines.

background stain to highlight the protein *bruchpilot* (*brp*), which is commonly used to identify regions in insect brain images (Figure 1.2). Details of the protocol used can be found in section 2.4

The confocal microscope was set up in dual scanning mode, with laser emission at wavelengths of 488nm and 546nm, and two corresponding capture channels. When EYFP expression was particularly strong sequential scanning mode was used to eliminate any cross emission from the EYFP channel to the background channel. No cross emission was observed from the background channel to the EYFP in either scanning mode. In most cases a 20x objective lens was used and whole brains were scanned at a resolution of approximately 1024x512x120 voxels, with a voxel size of approximately 0.55x0.55x1µm.

Figure 1.2 shows some typical scans showing EYFP expression in green and background expression in magenta.

Protein	Lines	D.E.	Protein	Lines	D.E.	Protein	Lines	D.E.
14-3-3epsilon	4	0	Dek	2	0	orb2	2	0
Abl	2	0	dlg1	3	0	ps	3	0
aop	5	0	Dys	2	1	Rab1	2	0
bel	2	1	ed	2	0	RnpS1	2	0
Bsg	3	1	Fas2	2	1	Rtnl1	2	0
Cam	2	0	Fas3	3	1	sdk	3	0
CBP	2	1	fax	2	1	Sema-2a	3	1
Ced-12	2	0	fl(2)d	2	0	sgg	3	0
CG10992	2	0	flw	2	0	Sh3beta	2	0
CG11486	3	0	gish	2	0	shot	2	0
CG14207	2	0	glec	2	0	Smr	2	0
CG1640	2	0	heph	3	0	Src64B	2	1
CG16812	2	0	Hrb98DE	3	0	stai	2	0
CG17839	5	2	hth	6	2	Su(Tpl) / Mi-2	9	1
CG18769	2	0	Imp	2	0	tau	2	0
CG2678	2	1	kst	2	0	Ten-m	2	0
CG31352	2	0	kuz	2	0	Tis11	2	0
CG32683	4	0	larp	2	1	Tm1	2	1
CG34339	2	0	Lk6	2	0	tmod	2	0
CG34422	2	0	Lsd-2	2	0	tral	2	0
CG4673	2	0	mbl	2	0	trol	2	0
CG9650	2	0	MTA1-like	2	0	zip	2	1
crp	2	0	Myo10A	3	1			
dally	2	0	N	2	0			
						Total	172	19

Table 3.2: Proteins trapped multiple times in the CPT project, with insertions occurring at different locations in the genome. Column “D.E.” shows the number of lines that show different expression to the most commonly observed expression pattern for that protein. Disagreement in expression patterns is observed in 11% of lines.

3.1.2 Protein Expression

It is important to note that the genetic modification of proteins by insertion of the artificial EYFP construct may affect the structure and therefore the localization of the trapped protein. Ideally this effect is small enough that the trapped protein can function as it normally would, however this is not always the case so results must be interpreted with care.

Most proteins of interest to neurobiology are transported away from the cell body to perform functions in other areas of the neuron such as in the dendrites and axons, or at the synapse or the axon initial segment (AIS). For *Drosophila* the anatomy of the brain is arranged with the neuron cell bodies located in a shell around the outside of the brain, which is referred to as the "brain cortex", (the area showing protein expression in green in Figure 1.2). Cell structures more relevant to neurobiology are located in the middle of the brain where axons and dendrites form a densely packed network of interconnected projections, which is collectively referred to as neuropil, and forms anatomical structures as shown by the *brp* antibody stain shown in magenta in Figure 1.2. This arrangement of the brain anatomy is convenient because proteins that are only observed in the brain cortex and not in the neuropil are likely to be less interesting for neurobiology. Trapped proteins that do not localize properly are observed in the brain cortex, so protein expression patterns observed in the neuropil are less likely to have been damaged by the insertion process, and are also more likely to be involved in important brain function.

The CPT lines also provide a useful set of data with which to assess the performance of the protein trap method. Some brain proteins captured by the screen have previously been imaged by other methods such as IHC, so expression patters can be compared to existing data. Also, some proteins were captured by several different lines in the CPT screen, at slightly different insertion points in the genome. This allows us to compare different insertions to each other. In most cases expression patterns observed are the same, however for some genes such as Fas2 we see one line (CPTI-000483) with an expression pattern matching the previously reported expression (Figure 3.4 B-D) and another line (CPTI-001279) showing expression only in the brain cortex (data not shown).

Overall, expression is consistent with previously reported results for 26 out of 31 protein trap lines where some pattern of expression is expected (table 3.1). For genes that are trapped more than once in the screen only 19 out of 172 protein trap lines show

disagreement between the different lines (table 3.2). These results are encouraging, and show that the protein trap method is reliable, but care should be taken when interpreting any individual result.

3.1.3 Results

It is not practical to discuss all proteins imaged by this screen, however to highlight the importance of proteomics research for neurobiology I will summarize expression patterns for some important genes here.

Behavior

dlg1 is a protein known to localize at the synapse and can be observed in the calyx of the mushroom body synaptic structures (Figure 3.2 A-D). *dlg1* has orthologues in mouse and humans and has been shown to be important for learning in fruit fly [128, 129, 64, 7]. There are 12 splice variants of *dlg1* listed on the FlyBase website (RA-RL). There are three protein trap lines in the BrainTrap database, with insertions at different locations in the genome. Two lines, CPTI-0002596 and CPTI-002860, trap all splice variants except for RC, RI, and RJ (Figure 3.2 A, B). The third line, CPTI-000207, traps all splice variants except for RA RD RE RG and RK (fig 3.2 C, D). These three lines show similar expression patterns and give us the possibility of investigating splice variation during development and localization at the anatomical and sub-cellular level.

nemy is required for learning in Drosophila [54, 46]. Trapped by line CPTI-004149, *nemy* shows expression throughout the brain, particularly in the optic lobes in a layer between the medulla and the lobula, as well as in the lateral horn and part of the mushroom bodies (Figure 3.5 A-C, and Figure 3.16).

Fas2 is a well known mushroom body marker, required for olfactory learning [128, 7, 69]. An antibody for this protein is widely used experimentally to highlight the mushroom bodies. Both *Fas2* and its vertebrate orthologue, *NCAM*, are involved in nervous system development [68]. Two lines trap this gene, and the expression pattern observed in CPTI-000483 shows the expected mushroom body staining (Figure 3.4, B-D), however line CPTI-001279 only shows low expression in the adult brain cortex (data not shown).

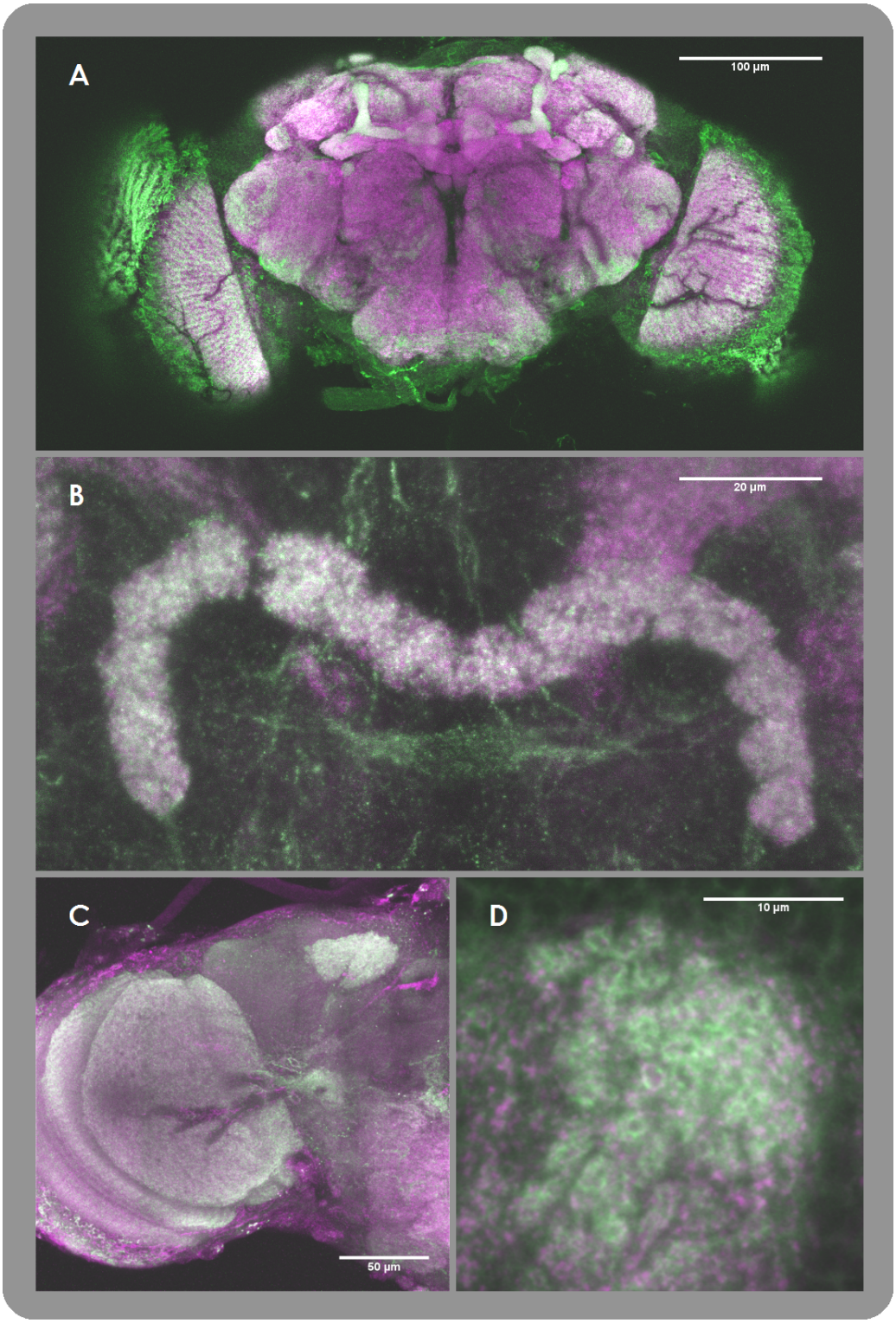


Figure 3.2: Maximum intensity projections (10μm: A,C. 5μm: B,D) of *dlg1* protein trap brain scans. Green indicates EYFP protein trap expression for lines CPTI-0002596 (A and B) and CPTI-000207 (C and D), both of which trap *dlg1*. Magenta indicates the anti-*brp* labeled, pre-synaptic background channel. Expression is higher in the optic lobes, mushroom bodies (A,C), protocerebral bridge (B), and localizes to synapses in the calyx of the mushroom body (D).

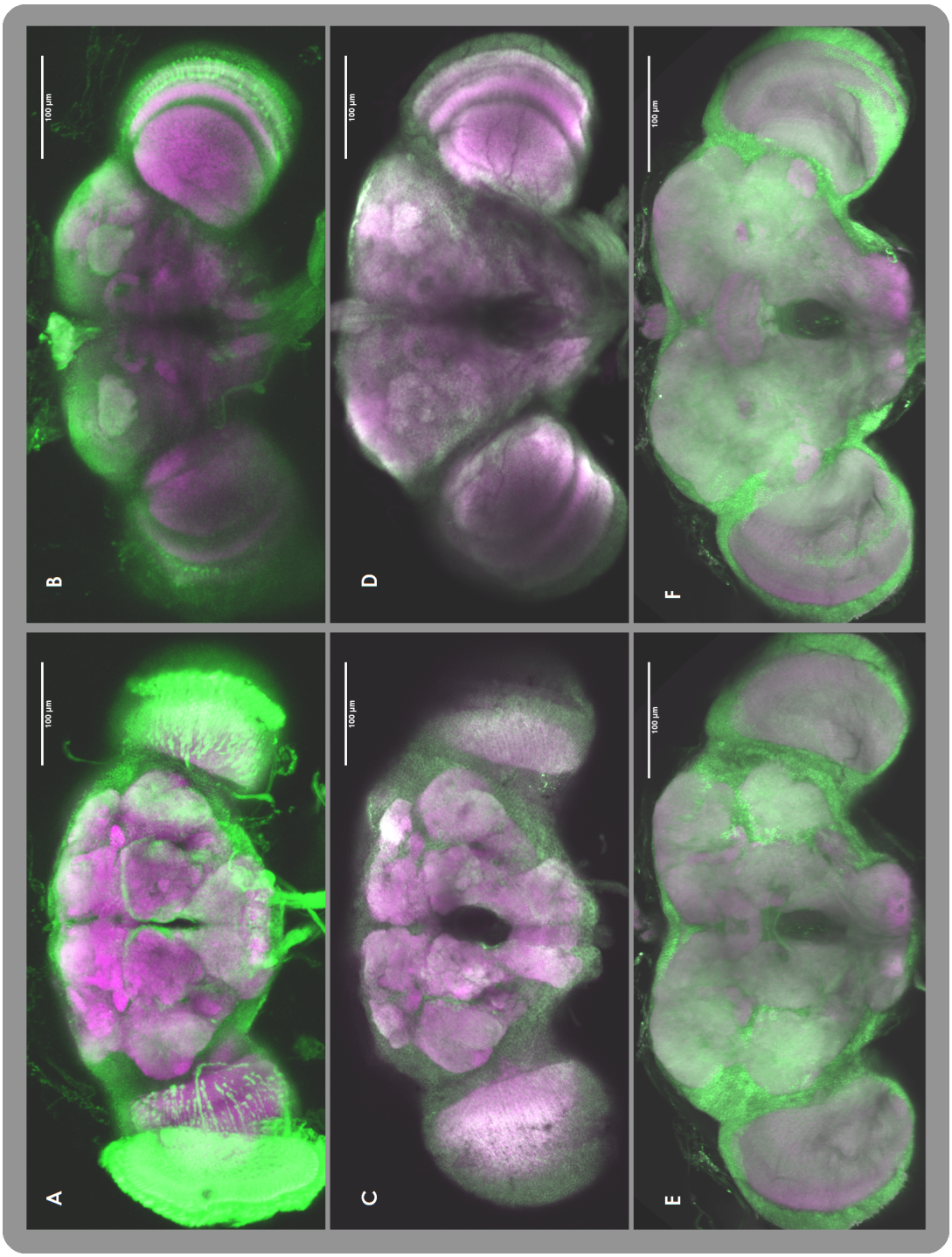


Figure 3.3: Maximum intensity projections (10μm) of *atpalpha*, *CaMKII* and *orb2* protein trap brain scans. Green indicates EYFP protein trap expression for lines CPTI-002761 / *atpalpha* (A and B) and CPTI-000994 / *CaMKII* (C and D), and CPTI-004204 / *orb2* (E and F). Magenta indicates the anti-*brp* labeled background channel. All images are maximum intensity projections from confocal image stacks available on the BrainTrap website.

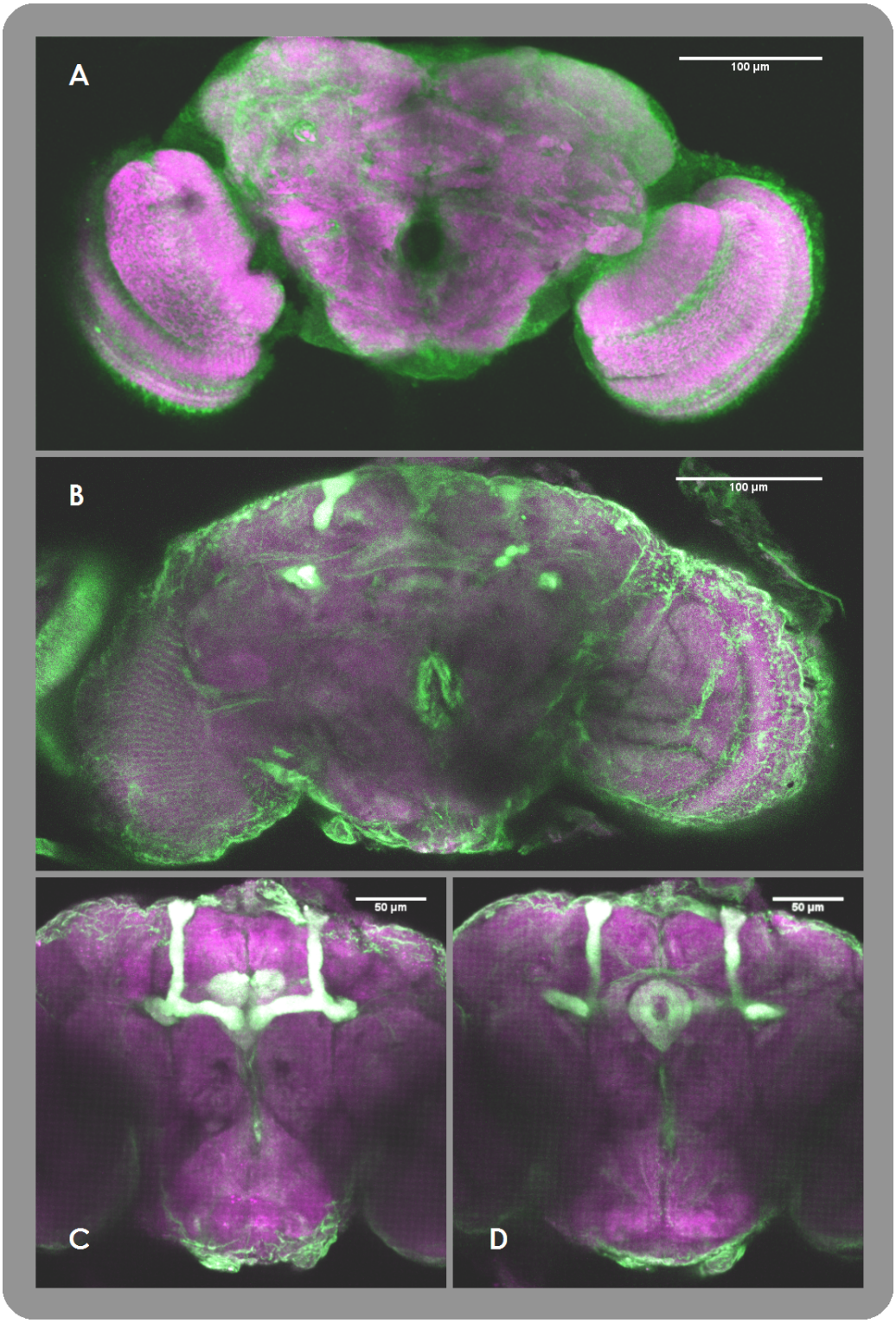


Figure 3.4: Maximum intensity projections (10μm) of *tau* and *Fas2* protein trap brain scans. Green indicates EYFP protein trap expression for lines CPTI-002745 / *tau* (A) and CPTI-000483 / *Fas2* (B-D). Magenta indicates the anti-*brp* labeled background channel.

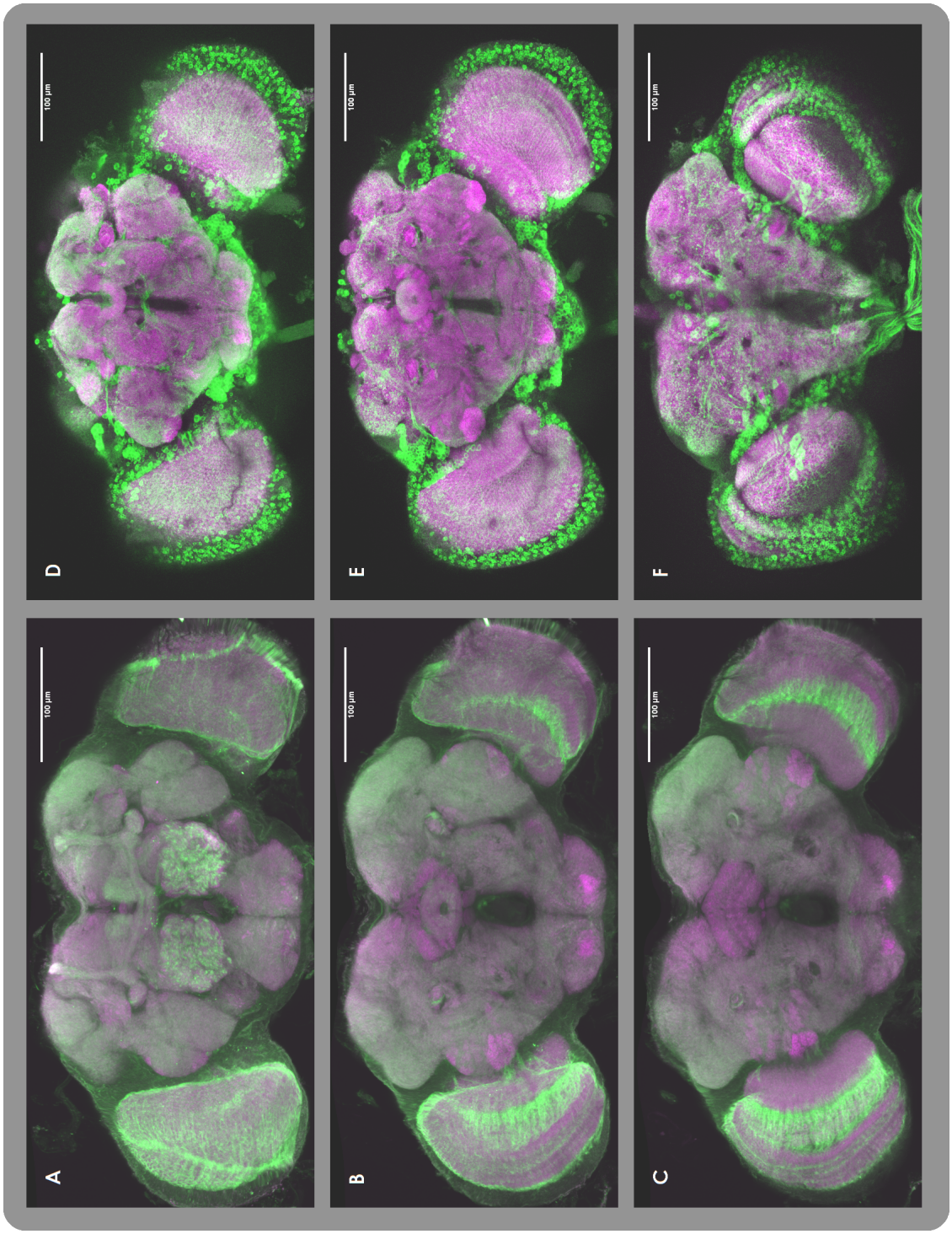


Figure 3.5: Maximum intensity projections (20μm) of *nemy* and *Gad1* protein trap brain scans. Green indicates EYFP protein trap expression for lines CPTI-004149 / *nemy* (A-C) and CPTI-000977 / *Gad1* (D-F). Magenta indicates the anti-*brp* labeled background channel. A-C: *nemy* shows expression throughout the brain, higher in the optic lobes, antennal lobes and the mushroom bodies. D-F: *Gad1* highlights inhibitory GABA neurons throughout the brain.

orb2 is required for long term courtship memory and synapse formation [58, 135, 83]. Trapped by line CPTI-004204, the protein shows expression throughout the brain, with higher expression in the lateral horn, and parts of the protocerebrum; areas known to be involved in courtship behavior [49, 60, 12] (Figure 3.3 E, F).

Neurobiology

Atpalpha (Atpα) is the alpha subunit of the sodium pump. This pump is fundamentally important for neuron function, it moves *Na+* ions out of the cell and *K+* ions into the cell to maintain the resting state of the cell, and has been used to study aspects of human disease in *Drosophila* [16, 67]. Protein trap line CPTI-002761 shows expression throughout the brain, elevated in areas such as the optic lobes and calyx of the mushroom body (Figure 3.3 A, B). Corresponding beta subunits of the sodium pump, *nrv1* and *nrv2* are also in the BrainTrap data.

CaMKII is an important and well studied signaling protein for synaptic plasticity which also regulates localization of already mentioned proteins *Fas2* and *dlg1* [64, 100, 7]. There is only one line in the screen that traps this protein, CPTI-000994 which shows expression throughout the brain, with higher levels in the optic lobes, and the calyx of the mushroom body (Figure 3.3 C, D, and Figure 3.15). The protein *Cam*, also trapped by lines CPTI-002888 and CPTI-001284, is known to interact with *CaMKII* in a well studied cascade that occurs after an action potential to regulate synaptic plasticity [30, 2].

Gad1 is a useful protein that highlights all neurons that express *GABA*, a the principle inhibitory neurotransmitter in the fruit fly brain [27, 82, 53]. Images show a subset of neurons activated (Figure 3.5 D-F), and some regions such as the ellipsoid body show concentrated expression, indicating that many inhibitory neurites are located here.

tau is a microtubule associated protein (MAP), homologue to the human *tau*. The protein is implicated in several neurodegenerative diseases known as tauopothies, including Alzheimer's disease, where neurofibulary tangles or plaques consisting of large amounts of *tau* form where neurons die; many aspects of these diseases have been studied in *Drosophila* [40, 137, 47, 17, 131, 116]. The protein trap lines show expression throughout the brain, higher in areas where many neurons projecting in the same direction form tracts (Figure 3.4 A).

3.1.4 BrainTrap Website

BrainTrap is part of a large scale collaborative research project to collect expression information for as many proteins as possible from protein-trap lines generated by the CPT project [113]. All image data generated by the brain imaging screen have been released on an online database, BrainTrap, which at time of publication contains 884 3D image stacks from 535 CPT lines. On the home page (Figure 3.6) several proteins of interest are highlighted with links directly to the interactive viewer.

3.1.4.1 Interactive Viewer

The main functionality provided by the BrainTrap website is to view the protein trap images in an interactive viewer, in a similar way to desktop applications for viewing microscopy images (Figure 3.7). The interactive viewer consists of a horizontal slider that can be used to navigate through the stack in the z direction, links to change image size (up to 100% of the original confocal scan) and buttons to select which channels are visible. To the right of the viewer is a list of annotations that have been made against that stack. Annotations linked to a location in the stack can be clicked to move the viewer to the corresponding z location and highlight the area of interest in the viewing window, as shown in Figure 3.7. In addition to the online viewer, confocal images files can be downloaded in the original format.

3.1.4.2 Annotation

Annotations have been made against the FlyBase Drosophila anatomical ontology [34]. Annotators can log in and add new annotations using the same interactive web interface with ontology term and synonym auto completion to help the annotator use the ontology even if they are not familiar it. Free text annotation and ontology terms can be combined to provide additional information. The provenance of all annotations is retained in the database and, where possible, annotations are linked to a 3D location in the volume showing the area of expression. Annotation data for each CPT line can be downloaded in extensible markup language (XML) format, or summarized in a gene comparison table for a subset of selected genes (Figure 3.10).

3.1.4.3 Search

BrainTrap takes advantage of the ontology structure used for annotations by providing a fast ontology driven search page. An anatomical search query can be specified on the

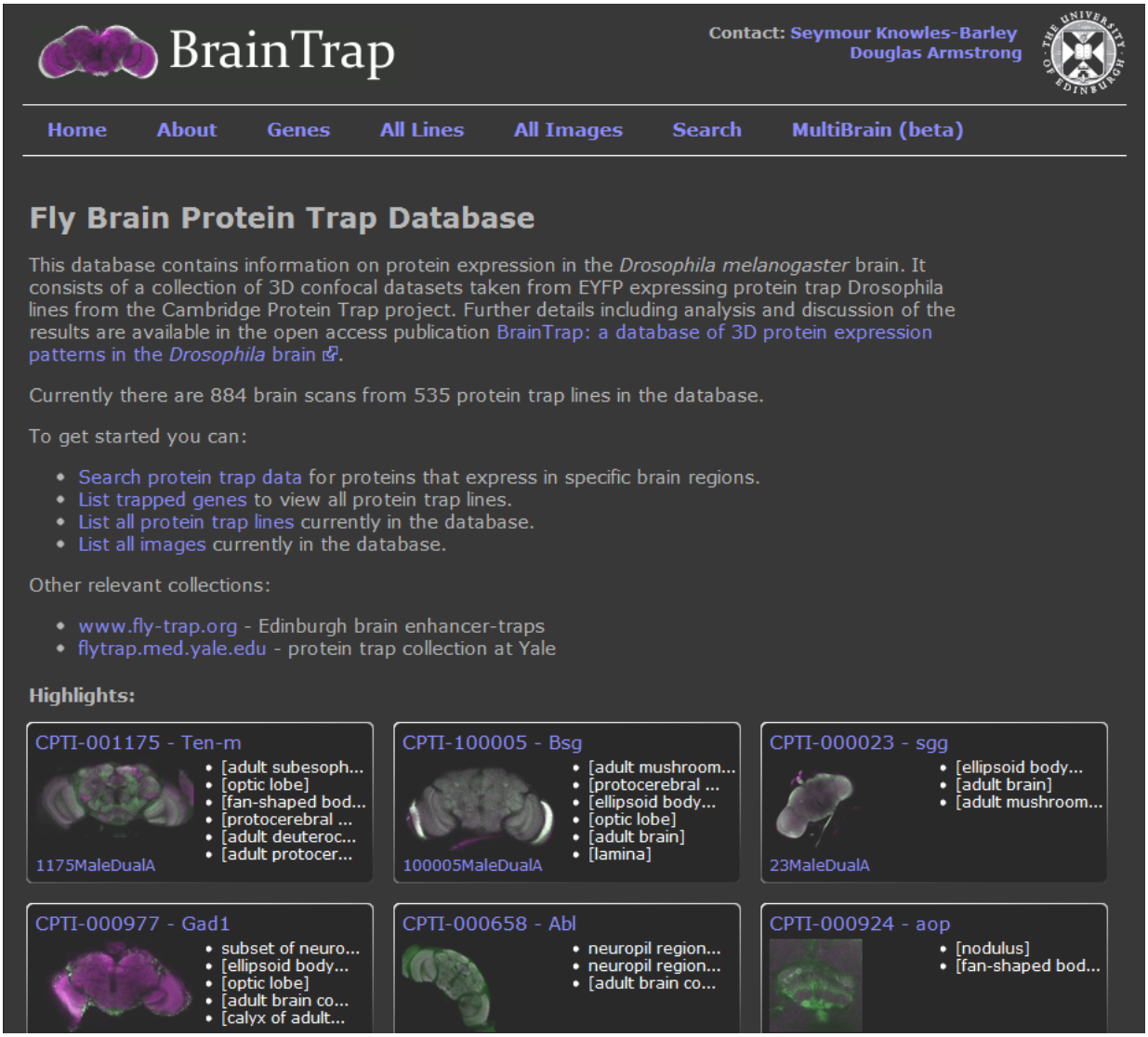


Figure 3.6: The BrainTrap home page displays information about the project and links to related websites. Thumbnail previews of some images are visible at the bottom of the page.

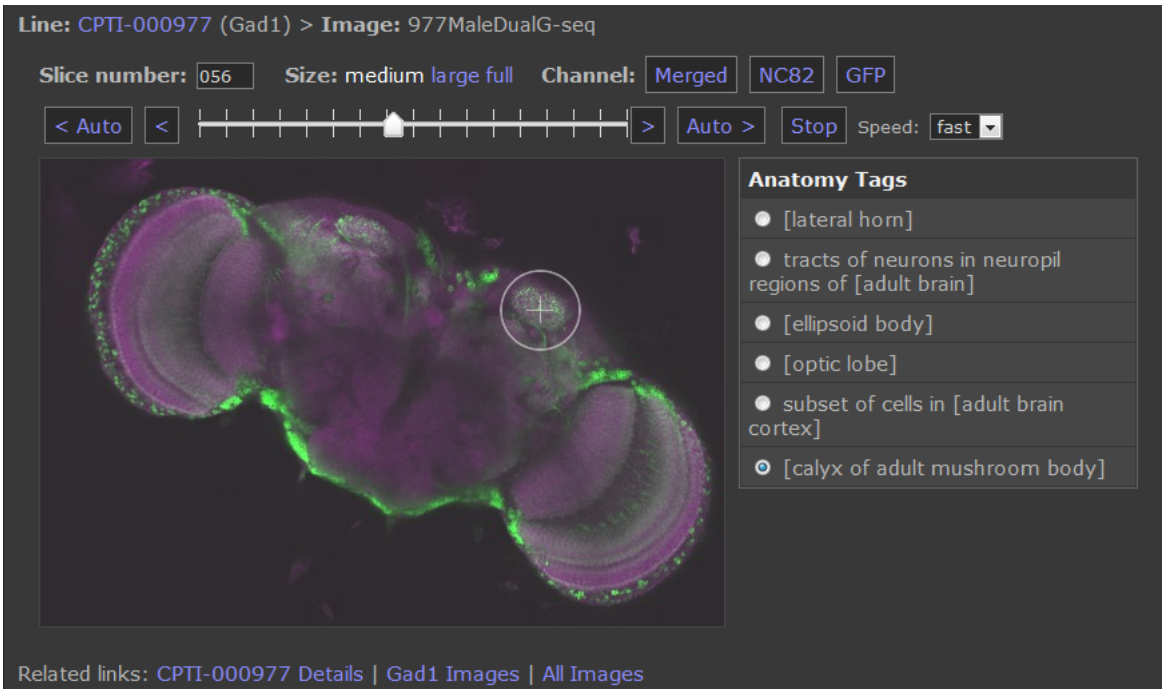


Figure 3.7: BrainTrap interactive viewer. Controls at the top of the screen allow the user to change the viewing size, up to 100% of the original confocal scan, select the visible channels, and navigate through the image stack. To the right of the screen annotations made for this stack are visible. Clicking on an annotation automatically highlights the area marked by the annotator, as shown above for the calyx of the mushroom body.

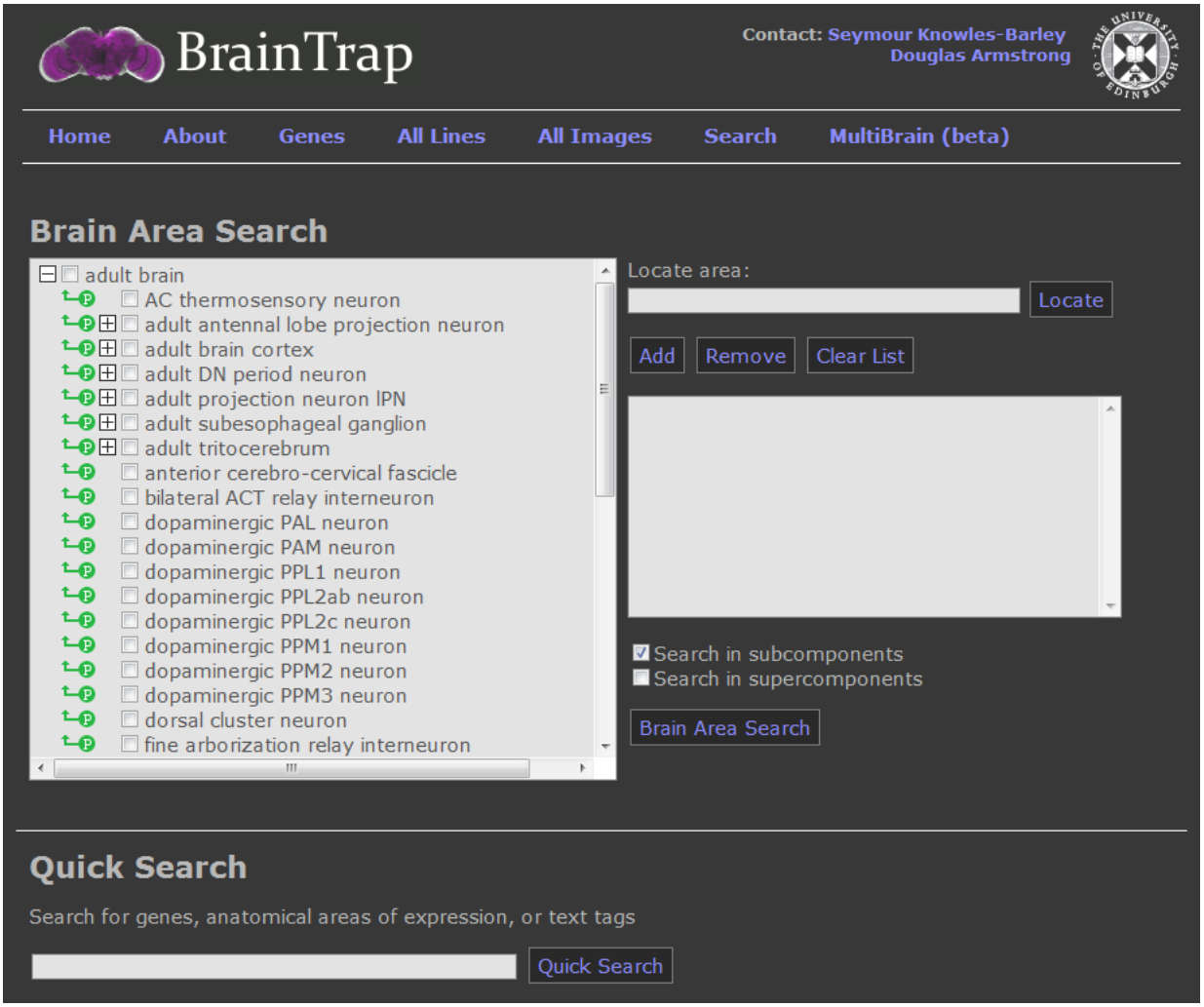


Figure 3.8: BrainTrap search form. On the left of the screen the FlyBase anatomical ontology can be explored, and on the right a text search box allows the user to search for anatomical terms. This form is used to specify an ontology-based anatomical search for protein expression patterns.

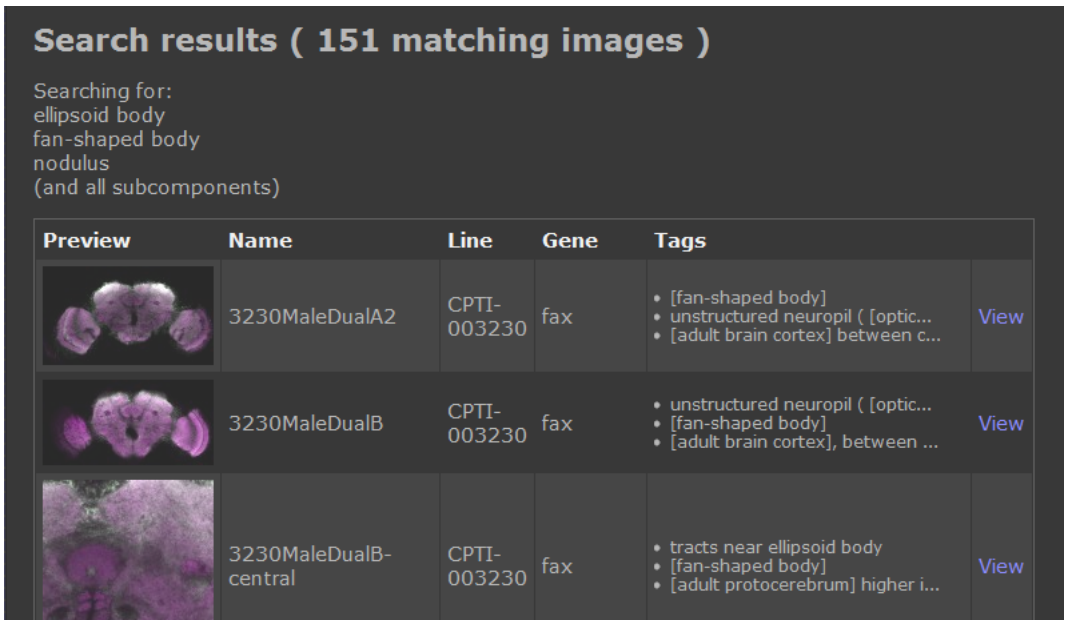


Figure 3.9: BrainTrap search results are displayed in summary format with annotation tags and thumbnail previews of brain scans.

web form shown in Figure 3.8. Areas can be easily specified by browsing the ontology tree (shown on the left) or by typing the area name or synonym into the search box. Anatomical areas are suggested as the user types and synonyms or old terms for the anatomical areas can be replaced with the correct term from the ontology. Once an area is selected the "Locate" button locates and highlights the area in the ontology tree at the left of the page. The subcomponents or supercomponents check boxes can be used to specify the type of search to perform. The default is to search for all images tagged with the selected term(s) or any subcomponents of those term(s), as determined by the anatomical ontology. For example, searching for "adult mushroom body" annotations will also match images tagged in a subcomponent, such as the "calyx of the adult mushroom body". Search results are shown in a preview format (Figure 3.8).

At the bottom of the page a "Quick Search" facility also allows you to search by text for gene names, or text in the annotations.

3.1.4.4 Other features

To help users navigate the website, 7 links appear in the top navigation bar (Figure 3.6). In addition to the Home and Search pages already discussed, there are 5 more areas of the website, and an administration area:

About Shows details of the methods used to generate CPT lines as well as the protocol

Listing all trapped genes

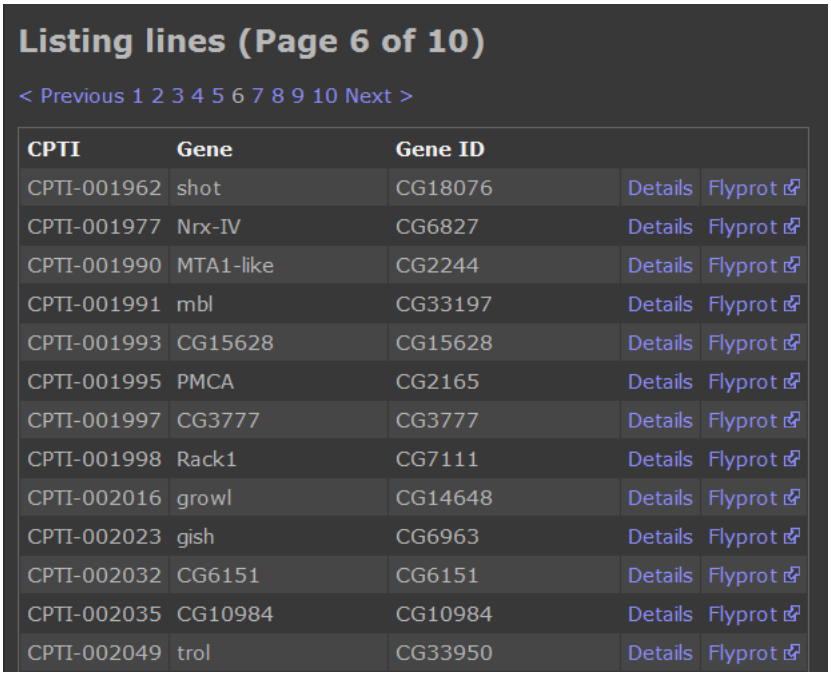
Compare Selected | Reset

	Gene name	Gene ID	Lines			
☐	???	???	69	Lines	Images	
☑	14-3-3epsilon	CG31196	4	Lines	Images	Flybase
☐	ab	CG4807	1	Lines	Images	Flybase
☑	Abl	CG4032	2	Lines	Images	Flybase
☐	ade3	CG31628	1	Lines	Images	Flybase
☐	ade5	CG3989	1	Lines	Images	Flybase
☐	AGO2	CG7439	1	Lines	Images	Flybase
☐	Ald	CG6058	1	Lines	Images	Flybase
☐	Alh	CG1070	1	Lines	Images	Flybase
☐	alpha-Cat	CG17947	1	Lines	Images	Flybase
☑	Amph	CG8604	1	Lines	Images	Flybase
☐	aop	CG3166	5	Lines	Images	Flybase
☑	Arc-p34	CG10954	1	Lines	Images	Flybase
☐	Argk	CG32031	1	Lines	Images	Flybase

Comparing trapped genes

	14-3-3epsilon	Abl	Amph	Arc-p34
adult antennal lobe	✓			
adult brain	✓			✓
adult brain cortex	✓	✓	✓	✓
adult deuterocerebrum		✓		
adult mushroom body	✓			
adult pedunculus	✓			
adult protocerebrum		✓		
adult subesophageal ganglion		✓		
ellipsoid body	✓			✓
fan-shaped body	✓	✓		
lamina			✓	✓
medulla				✓
mushroom body alpha-lobe	✓			
mushroom body gamma-lobe	✓			
optic lobe		✓		✓

Figure 3.10: The BrainTrap gene comparison feature allows expression patterns for multiple selected genes to be compared side by side.



Listing lines (Page 6 of 10)

< Previous 1 2 3 4 5 6 7 8 9 10 Next >

CPTI	Gene	Gene ID		
CPTI-001962	shot	CG18076	Details	Flyprot
CPTI-001977	Nrx-IV	CG6827	Details	Flyprot
CPTI-001990	MTA1-like	CG2244	Details	Flyprot
CPTI-001991	mbl	CG33197	Details	Flyprot
CPTI-001993	CG15628	CG15628	Details	Flyprot
CPTI-001995	PMCA	CG2165	Details	Flyprot
CPTI-001997	CG3777	CG3777	Details	Flyprot
CPTI-001998	Rack1	CG7111	Details	Flyprot
CPTI-002016	growl	CG14648	Details	Flyprot
CPTI-002023	gish	CG6963	Details	Flyprot
CPTI-002032	CG6151	CG6151	Details	Flyprot
CPTI-002035	CG10984	CG10984	Details	Flyprot
CPTI-002049	trol	CG33950	Details	Flyprot

Figure 3.11: The BrainTrap “line list” page displays all CPT lines available in the database, with links to further details in the BrainTrap site, or on the FlyProt site.

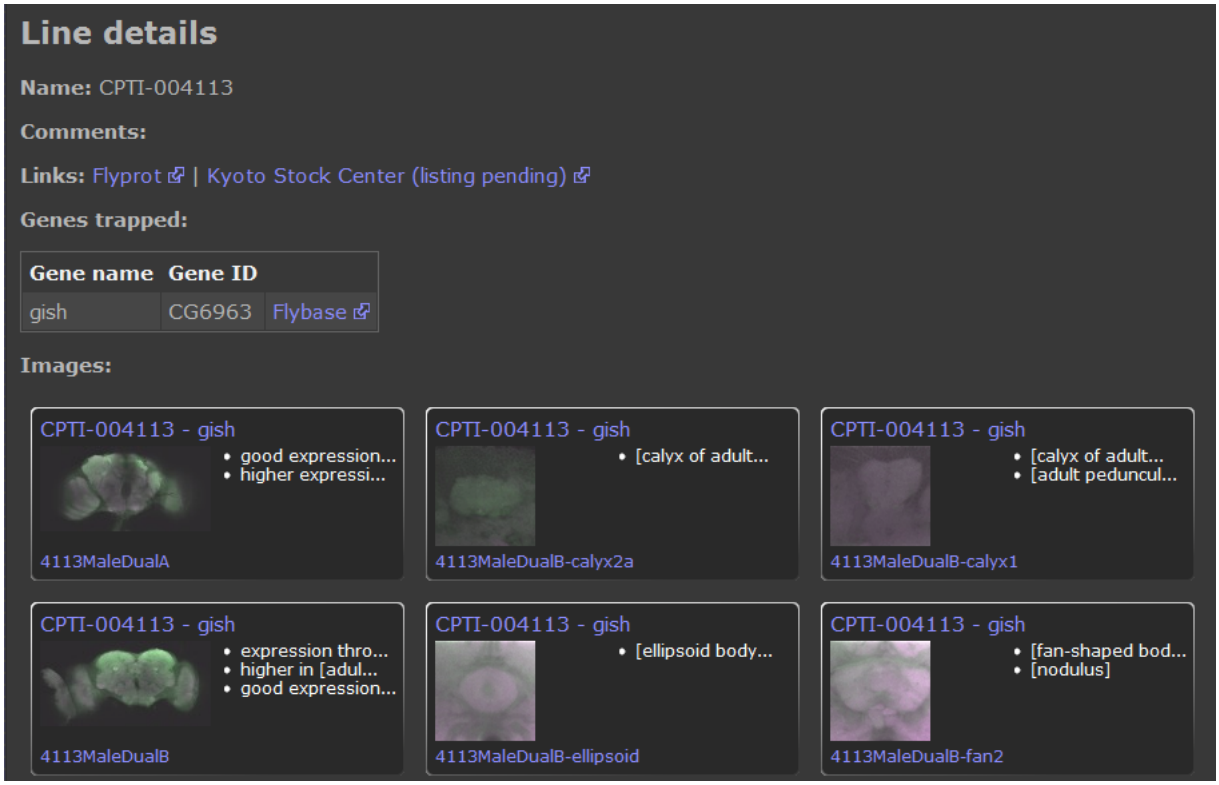


Figure 3.12: The line details page summarizes genes trapped by each line and provides links to the Kyoto Stock Center, the FlyProt website, and the FlyBase website. All confocal images are displayed in thumbnail preview format, with a summary of the annotations made for each image.

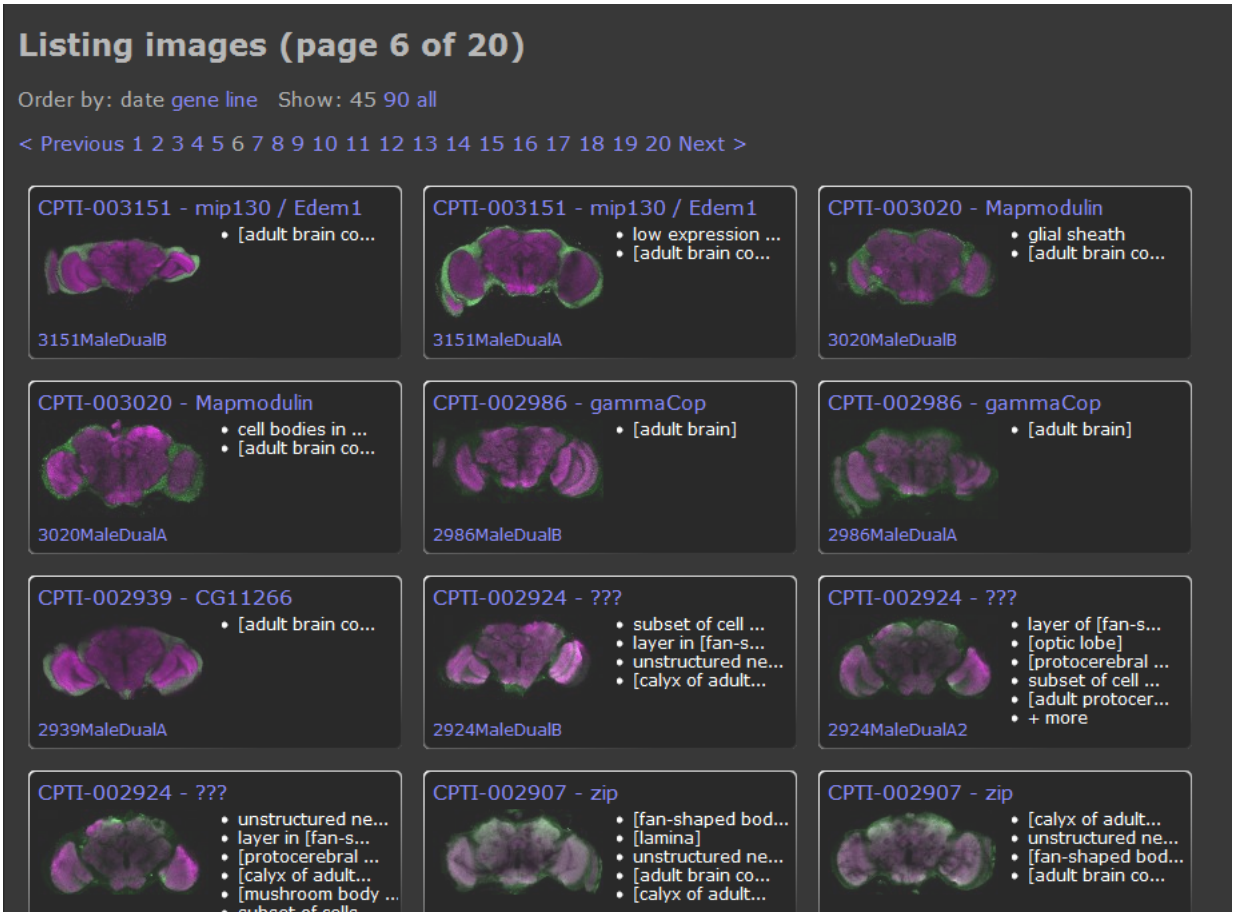


Figure 3.13: The BrainTrap “all images” list displays a summary of all images in thumbnail preview format, with a summary of the annotations made for each image.

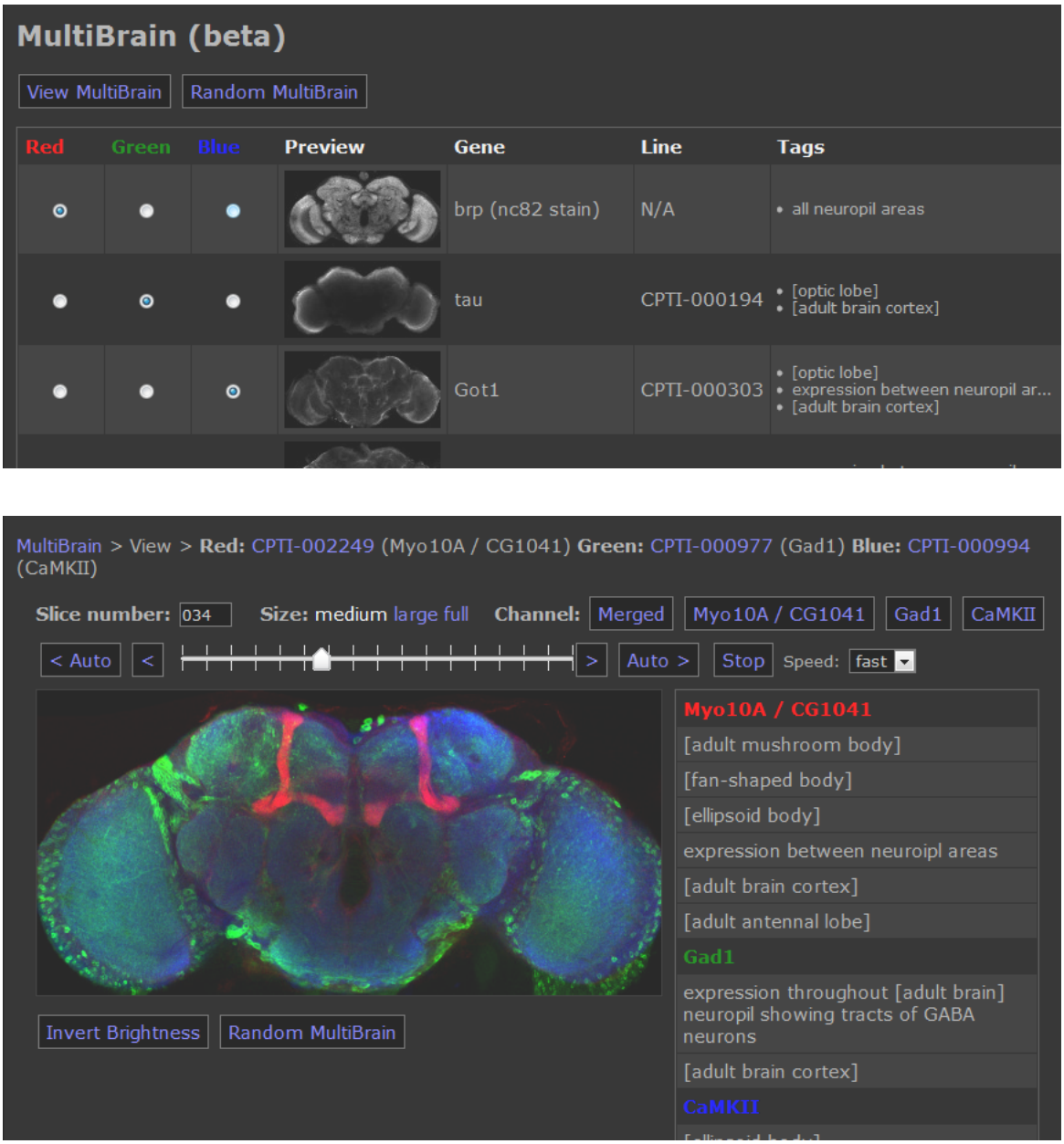


Figure 3.14: The BrainTrap MultiBrain page (top) allows the user to select up to three different expression patterns mapped to the same template brain. Clicking on “View MultiBrain” takes the user to the MultiBrain interactive viewer (bottom) where each expression pattern is displayed in a different colour, in the same viewing window. Stack viewing controls allow the user to change the size, the channels visible, and the position in the image stack. Annotations for each channel are visible on the right of the screen. The “Random MultiBrain” selects three expression patterns randomly in different colours.

used to generate images, and map images to a template.

Genes This page summarizes all the genes listed in the website (Figure 3.10, left). Links are provided to view all lines that trap each gene or to preview all images for the selected gene, with links to the interactive viewer. To compare gene expression patterns the user may simply select the genes of interest with check boxes at the right of the page, and click "Compare Selected". A table is displayed summarizing the annotations made for each gene (Figure 3.10, right).

All Lines List each CPT line with links to further details and to the Cambridge FlyProt site (Figure 3.11). The "further details" pages summarize each line, showing previews of all confocal brain scans in the database, with additional links to the Kyoto stock center where the lines can be ordered, and to the FlyBase listing for the gene(s) trapped by the insertion. Thumbnails of all the images available for the selected line are shown at the bottom of the page (Figure 3.12).

All Images Lists all images in preview mode, so that thumbnails of many images can be viewed at once. Links take the user to the interactive viewer directly (Figure 3.13).

MultiBrain On this page a demonstration of the template mapping results are available. Currently, 40 expression patterns have been registered against a template brain and this section of the site can be used to produce a composite image where each protein is visible in a different colour. On the MultiBrain page the user selects a colour (red, green or blue) for up to three different expression patterns (Figure 3.14, top). Then clicking "View MultiBrain" displays the interactive viewer with the three expression patterns combined into a single stack, each in the selected colour (Figure 3.14, bottom).

Admin Administrators can log in to add or modify annotations and comments online.

The BrainTrap website provides a valuable resource for any researchers to find proteins of interest and view their expression patterns in the fruit fly brain online. All annotations are available in machine-readable XML format to enable transparent data sharing throughout the scientific community. Data from the BrainTrap database, in conjunction with publicly available data from multiple sources, is used in Chapter 4 to link the levels of molecular biology, anatomy and connectivity. The protein trap lines recorded in BrainTrap can also be used to investigate the sub-cellular localization of

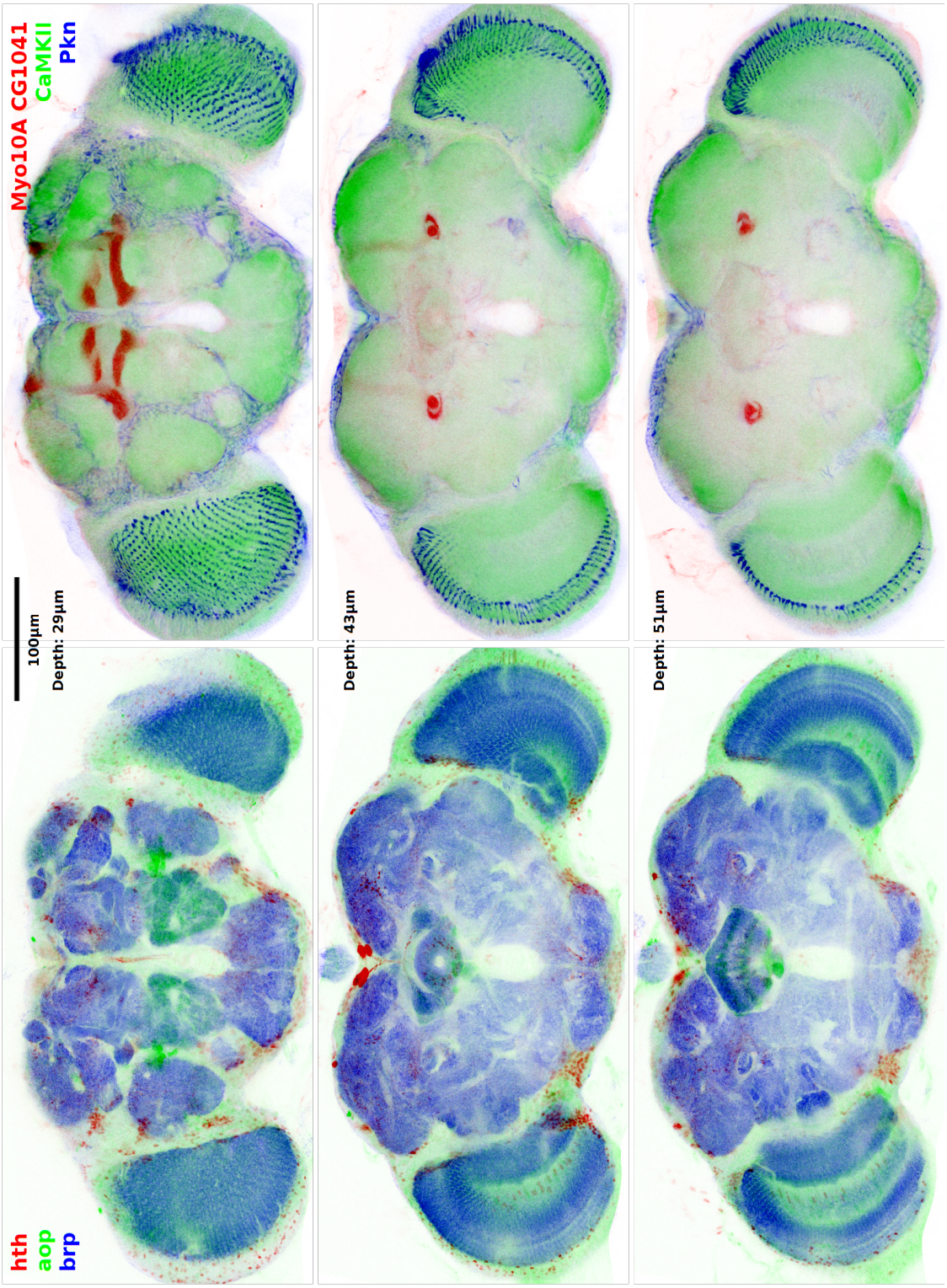


Figure 3.15: BrainTrap MultiBrain example images displayed on a white background, with darker regions indicating higher levels of expression. Each brain shows three expression patterns at once, mapped to the same template brain. Top: Expression from *Myo10A* / CG1041 (red), *CaMKII* (green) and *Pkn* (blue). Bottom: Expression from *hth* (red), *aop* (green) and the anti-*brp* background stain (blue).

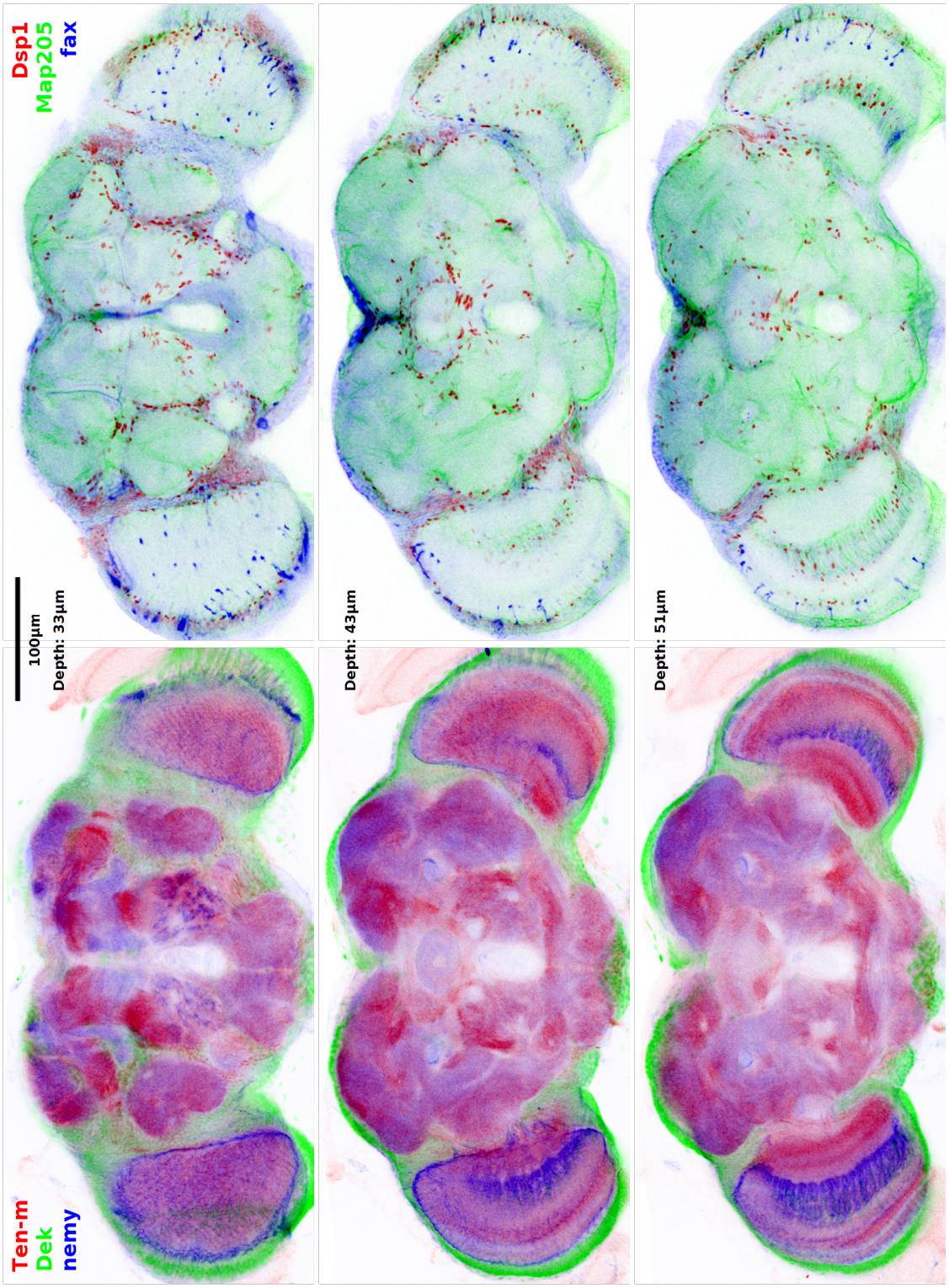


Figure 3.16: BrainTrap MultiBrain example images displayed on a white background, with darker regions indicating higher levels of expression. Each brain shows three expression patterns at once, mapped to the same template brain. Top: Expression from *Dsp1* (red), *Map205* (green) and *fax* (blue). Bottom: Expression from *Ten-m* (red), *Dek* (green) and *nemy* (blue).

proteins, including synaptic localization (Figure 3.2, D). In the next section I describe attempts made to image protein localization, in the same protein trap lines, at a higher resolution using electron microscopy.

3.2 Electron Microscopy

Some of the proteins imaged in the CPT lines are known to locate at the synapse, and play an important role in learning and memory, as discussed in Section 3.1.3. To investigate some of these proteins further we decided to investigate whether the protein trap lines could be detected with IHC at EM resolution. Immuno-staining at the electron microscope level is difficult to achieve, and sample preparation methods must be optimized for each protein and antibody used. The protein trap lines present a possible solution to this difficulty, where a single protocol could be identified for the EYFP insertion and this could be used to image any of the many protein trap lines modified in this way. The GFP family is widely used as a fluorescent marker in biology and a standardized protocol for imaging the molecule at the EM level would be very useful.

To learn more about electron microscopy imaging and to investigate immuno-staining of the GFP family of molecules under EM fixation conditions I traveled to Dalhousie University in Halifax, Nova Scotia, to collaborate with Ian Meinertzhagen and his lab.

While there I attempted several preparation methods to stain the EYFP insertion. In some cases weak staining was observed, but no clear results were obtained in the limited time I had for my visit. In the following sections I describe the experiments performed and results obtained.

3.2.1 Requirements

For the synaptic protein lines it is desirable to image an entire synapse, including the surrounding volume, to determine where proteins are in relation to the synapse. This would allow identification of pre- or post- synaptic proteins, measurement of how far away from the synapse proteins are located, and identification any preference for where in the synapse they localize. To image a small 3D volume such as this we can use serial section transmission electron microscopy (ssTEM). Very good EM preservation is required for this method to work well, and to observe the fine structure of a synapse

in three dimensions. Four major steps are usually used to prepare samples for this imaging:

1. Strong fixatives are applied to the tissue to cross-link the proteins together to preserve ultrastructure
2. Osmium is applied to cross-link the lipid membrane to preserve the ultrastructure and also to provide an electron dense stain for imaging.
3. The tissue is embedded in epoxy resin and heated to polymerize to a solid block
4. The polymerized block is cut into thin sections, usually 50nm thick, ready for TEM imaging.

Each of these three steps can damage protein immunoreactivity, depending on the fragility and reactivity of the protein and the antibody used to tag it.

There are two main options for the process of staining proteins at EM level:

1. Preembedding, where antibodies are applied after the first fixation step
2. Postembedding, where antibodies are applied after sections are cut.

In order to investigate the staining of EYFP at the EM level we attempted several methods for fixation, preembedding and postembedding, which are explained in detail in the following sections. Below is a summary of the methods attempted:

- Antibody embedding: Preembedding / postembedding
- Antibody: normal anti-GFP / Heat-denatured GFP antibody
- Area of interest: Mushroom body / central complex / lamina / optic lobe

3.2.2 Fixation

To achieve good image quality in electron microscope images, strong chemical fixatives such as paraformaldehyde (PFA), gluteraldehyde (GA) and acrolein (ACR) are usually used to preserve tissue structure, and osmium tetroxide is used to preserve and stain the lipid membrane of the cell walls. These fixatives can damage proteins to the point where antibodies no longer bind to their targets. Performing IHC in tissues fixed for EM imaging requires a balance between the level of fixation and the immuno-reactivity of the protein of interest. This balance is different for each protein,

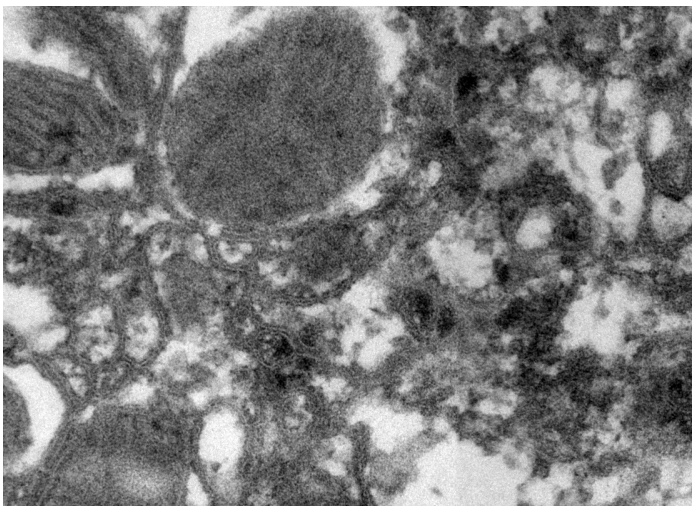


Figure 3.17: Preembedding with silver enhancement staining. Dark black areas show some staining, however in this case light fixation conditions caused poor ultrastructure preservation.

with some proteins preserving well even under strong fixative conditions and others being damaged easily. Unfortunately, both the fluorescence and immunoreactivity of the GFP protein family is easily damaged by fixatives, so a major part of this investigation is finding the right fixation procedure to preserve GFP immunoreactivity and tissue ultrastructure.

In an investigation into fixation methods I tried combinations of different fixative chemicals (PFA+GA or PFA+ACR), different concentrations of the fixatives, and performed the fixation over different lengths of time. The general protocols used for both pre- and post- embedding are described in Section 2.3.

3.2.3 Preembedding

For the preembedding method antibody staining is performed after aldehyde fixation. In this method no osmium staining is used. Instead a silver nitrate reaction is used to create electron opaque stain on tissue and antibody-attached gold particles. The increased amount of silver present in the sample can complicate the cutting procedure. No staining was shown at normal EM fixation levels, and at lighter fixation levels some staining was observed, however the tissue preservation was poor (Figure 3.17).

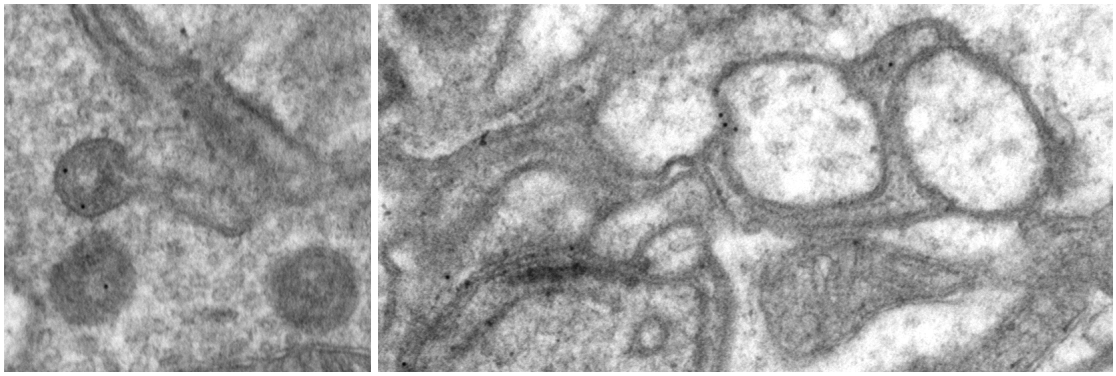


Figure 3.18: Postembedding with 10nm gold particles. *Bsg* protein trap line shows anti-GFP immuno-reactivity in the lamina at capitate projections and near membrane, however signal is weak. Image widths are 0.8μm (left) and 1.5μm (right).

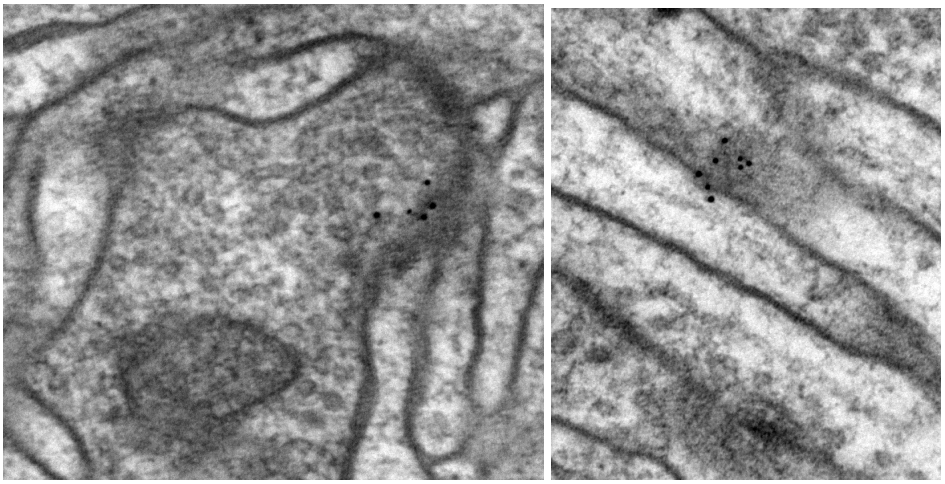


Figure 3.19: Postembedding with 10nm gold particles using heat denatured anti-GFP antibodies. Some staining was observed for *dlg1* protein trap lines in the calyx of the mushroom body (left) and for Fas2 in the stalk of the mushroom body (right), however signal is weak and possibly non-specific. Image widths are 1.1μm (left) and 0.8μm (right).

3.2.4 Postembedding

The main alternative method for immunostaining is postembedding. In this method the tissue fixation steps are performed first. Tissue is embedded and cut, then immunostaining is performed on the 50nm slices before imaging. High fixation levels were attempted first, to observe good preservation and osmium staining, making cutting and serial section imaging easier. After cutting, the embedding medium is etched away using acid, and antibodies are applied to the surface of the etched sample to attach to the exposed proteins.

Some positive results were observed for postembedding, showing some expression for *Bsg* near membrane, synapse and capitate projections in the lamina (fig 3.18). *Bsg* fluorescence in the lamina is very high, but the level of staining achieved under EM was only low, so further work is required to perfect this protocol.

One of the main problems encountered was that the harsh fixation conditions required for tissue preservation for EM quenched both EYFP fluorescence and immunoreactivity (Figure 3.20). A balance between fixation conditions and protein preservation must be found for this procedure to work properly. Another alternative that was attempted was to use heat-denatured GFP antibody (Figure 3.19). This antibody provided some weak staining, but possible non-specific staining was observed and results were inconclusive.

3.2.5 Alternatives

With no clear successes at the EM level I then investigated fluorescence and immunoreactivity under different fixation conditions at the light level. As shown in Figure 3.20, the EYFP fluorescence and immuno-reactivity is present after 1 hour of fixation, but is quenched after 2 hours of fixation. The mild fixation was not sufficient to achieve good preservation at EM level.

When investigations were carried out it was not clear how to continue and investigation of all possible alternatives for sample preparation would have taken a very long time. Possibilities for future investigation include:

Staining HRP / DAB reaction by oxidation. Commonly used to stain samples at the EM level, this reaction precipitates an electron dense stain. This method still requires preembedding of antibodies conjugated to HRP at the preembedding stage.

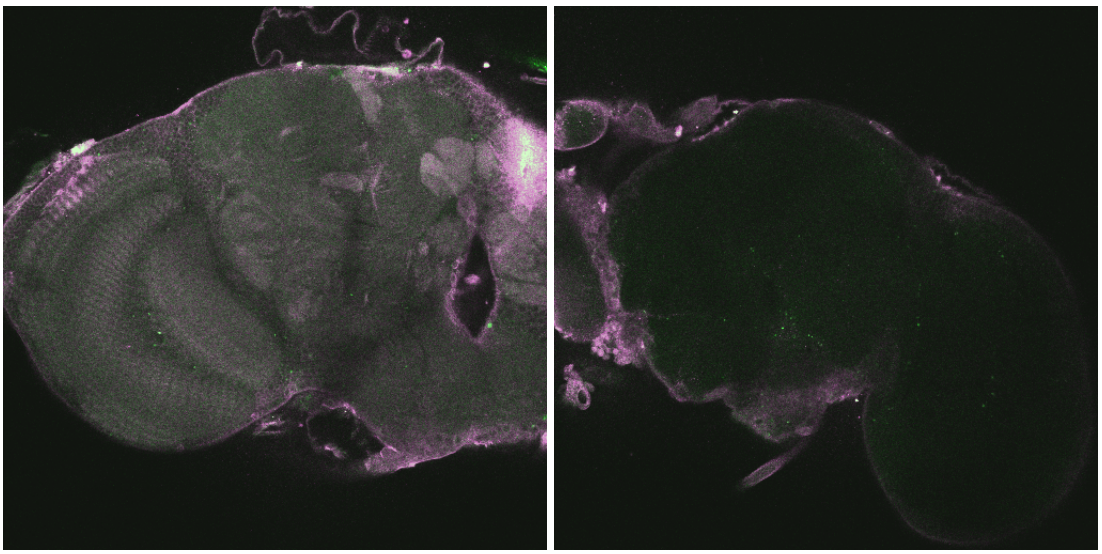


Figure 3.20: Immunoreactivity and EYFP fluorescence is quenched by EM fixation. (Left) Light fixation for 1 hour under 4% paraformaldehyde and 1% Gluteraldehyde. Both fluorescence (green channel) and immuno-staining (magenta channel) are present in the image. (Right) Normal fixation for 2 hours under the same conditions. In vivo fluorescence and immuno-staining are no longer present under these fixation conditions. Both images are 368.55μm wide.

Fixation — High pressure freezing / freeze substitution. In this method samples are first frozen under high pressure, and then fixation occurs very slowly over the following days. This method produces very good tissue preservation under mild fixation conditions, but requires specialist equipment.

Microwave fixation. For this method milder fixatives are combined with microwave excitation to speed up the fixation process.

Embedding — It is also possible to use a UV polymerized embedding medium such as LR White or Lowicryl rather than the heat polymerized epoxy resin, however this may make the cutting step more difficult.

Due to the time consuming nature of immuno-EM protocol development, further investigation was abandoned until more reliable methods of staining could be identified.

Since investigations were carried out, other labs have developed alternative protocols for staining proteins at the EM level and recent publications show that co-expression of EGFP and horseradish peroxidase (HRP) is effective at both light and EM level [72], and a new protein, miniSOG, was developed for the combination of

light and EM imaging [120]. Utilizing these techniques for the CPT lines would require additional modification of the genetic insertion to a new protein. P-element tags in the CPT lines allow straightforward modification of the existing construct in the same location. This provides at least two possible methods for staining protein trap lines at the EM level. Large scale EM screen of protein trap lines may now be possible using these techniques.

3.3 Summary

BrainTrap allows fast access to a large set of information; there are currently 884 3D image stacks from 535 protein trap lines available on the website. A wide variety of expression patterns are observed within these images. Areas or proteins of interest can be identified quickly with a simple search query and desired images can be accessed in a few mouse clicks. Large scale biological screens, such as the CPT Project, create huge amounts of raw data. Fast access to targeted information within this raw data is very important for the field of biological and brain science as large scale screens continue to grow in size. Structured annotation of the raw information also allows further analysis to extract useful information from the large data sets. The BrainTrap database will continue to grow if more protein trap data is made available, providing a valuable resource for *Drosophila* neurobiology and brain research in the future. The MultiBrain area of the site allows expression patterns to be compared directly between different protein-trap lines.

Investigations into immuno-labelling of the protein trap lines at the EM level could not be perfected enough to reliably image proteins in this way. There are several options for future research in this area; methods using alternative EM preparation or further transgenic techniques offer potential paths for establishing an effective protocol.

In the next chapter I investigate methods for using the protein expression data described in this chapter to link molecular biology with the levels of anatomy, and connectivity to provide further insight into each level, and to gain a more wholistic understanding of brain function.

Chapter 4

Network Analysis

The network is a useful mathematical abstraction common to the levels of brain understanding considered in this thesis. For example, genes are linked together in a network to regulate gene and protein production, proteins are linked in an interaction network, anatomical regions are linked by an ontology network, and neurons are linked together by a connectivity network.

The network representation allows us to apply mathematical analysis tools to these networks. In this chapter, I will use some of these tools to analyze the protein expression information from the BrainTrap database in combination with other networks to demonstrate some of the possible uses of this data. In Section 4.1, I use the BrainTrap data to improve protein interaction networks by ruling out many potential protein-protein interactions for the adult *Drosophila* brain. In Section 4.2, I produce a natural anatomical segmentation of the fruit fly brain based on a subset of protein expression patterns. This anatomical breakdown informs us about the similarities between regions of the fruit fly brain at the molecular level. And in Section 4.3, I show that when combined with homology data, the BrainTrap data can uncover similarities between insect and mammalian neurobiology, and how evolution has driven brain development.

Traditionally, gene and protein networks are considered separately to neural connectivity networks because of the lack of data connecting the two areas of study and the difficulty of collecting both types of data at the same time. Anatomy provides a 3D environment in which both protein expression and neuronal connectivity networks are embodied. This provides a bridge with which we can join the levels of molecular biology and connectivity, and while a full neural connectivity network is not yet available for *Drosophila* we can begin to look at how this data can be joined and analyzed. Motif analysis is a popular way of analyzing networks to distill out smaller building

blocks of the network and compare the building blocks of different types of networks for similarities. In Section 4.4.1, I describe an extension to the motif analysis method that considers connectivity networks in the context of their biological properties, such as protein expression or neuron type information. This analysis method allows a more detailed understanding of the underlying network properties than motif analysis alone.

4.1 Protein Interaction Networks

Large protein interaction databases are available online, for example the DroID database [142] contains protein-protein interactions for Drosophila proteins collected from many data sources. Most of the data in these interaction databases is obtained using the yeast two-hybrid method. This method involves expressing the two proteins of interest inside a yeast cell, with two halves of a transcription factor bound to them. If the proteins interact the transcription factor is activated and a reporter gene is expressed. Expression of this reporter gene is measured to give an indication of the level of interaction between the two proteins. This method can be scaled up to investigate many thousands of protein-protein interactions [29, 132, 102]. Unfortunately this method produces many interaction results that are not observed *in vivo*. Due to protein expression in different anatomical locations, or different stages of development, these interactions may not ever happen *in vivo*.

Using data from the BrainTrap website we can determine which proteins are expressed in the same location in the brain, and which proteins do not localize to the same place in the adult brain of *Drosophila*. Any protein-protein interactions detected in the yeast two hybrid experiments that do not express in the same location can be ruled out as unimportant for the adult *Drosophila* brain. Note that this does not mean that the interaction does not occur at all (the interaction may occur in a different organ, or at a different developmental stage) but it does tell us that the interaction is not important in the adult brain.

Figure 4.1 shows some of the interactions listed in DroID that can be ruled out by the BrainTrap data. The DroID database also calculates a confidence for each interaction as described in [141]. The proteins shown in Figure 4.1 have many other interactions that are not ruled out, and the interactions that are ruled out have low confidence scores in the DroID database. Table 4.1 summarizes all the protein-protein interactions that are called into question in this way. An additional 107 interactions between the protein Ubx (a developmental gene) and other proteins are not listed.

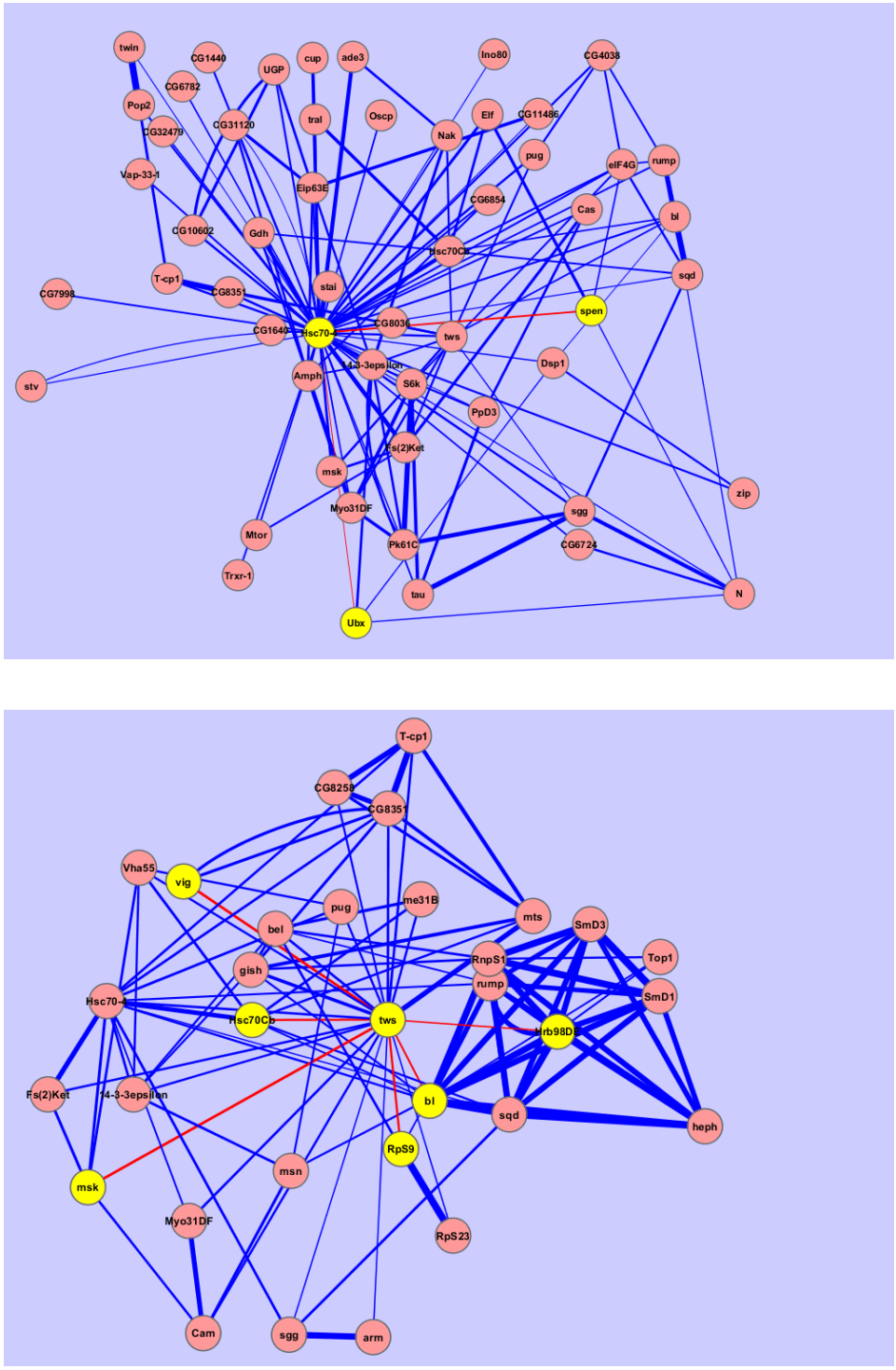


Figure 4.1: Interactions between proteins that can be ruled out by protein expression information in the BrainTrap database. Circles indicate a subset of proteins present in the BrainTrap database. Lines indicate protein-protein interactions from the DroID database. Thickness of the lines indicate the confidence scores for each interaction as listed in DroID. Each figure highlights the interactions that can be ruled out for two central proteins *Hsc70-4* (top) and *tws* (bottom). Interactions highlighted in red between proteins highlighted in yellow are potentially ruled out by the BrainTrap data.

Protein 1	Protein 2	Confidence	Protein 1	Protein 2	Confidence
Abl	Nrg	0.34	if	msk	
aop	spen		if	Nrg	0.403
arm	spen		if	Ten-m	
bl	tws	0.2	JIL-1	Nrg	
bmm	Dsp1		kst	Nrg	
bmm	Lsd-2		me31B	tws	0.243
CG30084	if		mge	mip130	
CG6983	Dsp1		mip130	nemy	
CG8258	tws	0.219	mip130	Nrg	
dally	Dsp1		mip130	tws	
dap	spen		mip130	uzip	
dlg1	Nrg		MTA1-like	spen	0.372
Dsp1	Nrg		N	spen	
Dsp1	tws		nemy	pbl	
ed	Nrg	0.418	Nrg	Nrx-IV	
eIF4G	spen	0.184	Nrg	shot	0.498
Elf	spen	0.356	polo	spen	0.349
Fas2	Nrg		sm	tws	
gish	spen	0.29	Smr	spen	0.411
Hsc70-4	spen	0.218	spen	Vha55	0.208
if	mew		T-cp1	tws	0.26
if	mip130				

Table 4.1: Interactions ruled out by BrainTrap expression data. Each protein-protein interaction is listed only once, with the first protein alphabetically appearing in the first column. The confidence column indicates the protein-protein interaction confidence score calculated from experimental and computational evidence as listed in the DroID database [142, 141]. Scores are between 0 and 1. An additional 107 interactions between the protein *Ubx* (involved in development) and other proteins are not listed.

Further investigations have been carried out using the CPT lines to investigate protein-protein interactions using protein affinity and mass spectrometry [107]. Combined with the localization data, this shows that the protein trap method is very useful for improving our knowledge of protein interaction networks.

4.2 Natural Segmentation

By using image data mapped to a template brain, it was also possible to cluster areas of the brain together based on the pattern of protein expression found at each point in the brain. This is similar to the approach taken in [9] of clustering data from the ABA using k-means to identify different anatomical areas in the mouse brain based on the gene expression profiles at each location.

Unfortunately the quantity of data available for *Drosophila* does not match the amount available in the ABA, which allowed the analysis in [9], to segment brain regions based on 3041 gene expression patterns mapped to the same template. Here I demonstrate the same techniques based on 18 protein expression patterns, selected from the 40 expression patterns mapped to a template brain using CMTK. Sparse expression patterns were not used, and where several very similar expression patterns existed, only one representative pattern was chosen.

For *Drosophila* brain segmentation, a subset of between 10 and 18 brain images im_{1-n} mapped to the same template brain (as described in section 2.4.2) were selected. Signal channels were first smoothed using a 5x5 pixel 2D Gaussian smoothing kernel, with a standard deviation of size 1 pixel, and then down-sampled to a size of 262x150x94 pixels. Each image pixel was then represented as an n-dimensional vector, $v_{x,y,z} = [p_{(1,x,y,z)}, p_{(2,x,y,z)}, \cdots, p_{(n,x,y,z)}]$, where $p_{(i,x,y,z)}$ represents the pixel at location (x,y,z) in image i. Vectors of Euclidean norm length $||v||$ less than 100 were assumed to be background pixels and not used for clustering calculations. Seed points were selected either randomly, or manually from 15-25 large anatomical structures in the *Drosophila* brain. Each manual seed point consisted of only one pixel from the middle of the anatomical structure, and only from one side of the brain. Clustering methods were then used to identify the same number of clusters as seed points, based on expression vectors v for all non-background pixels. Classification was performed on full-sized 1048x600x94, n-dimensional images. Note that the spacial location of each pixel was not considered for the clustering.

K-means results, for $n = 10$ with manually selected seed points and 15 clusters

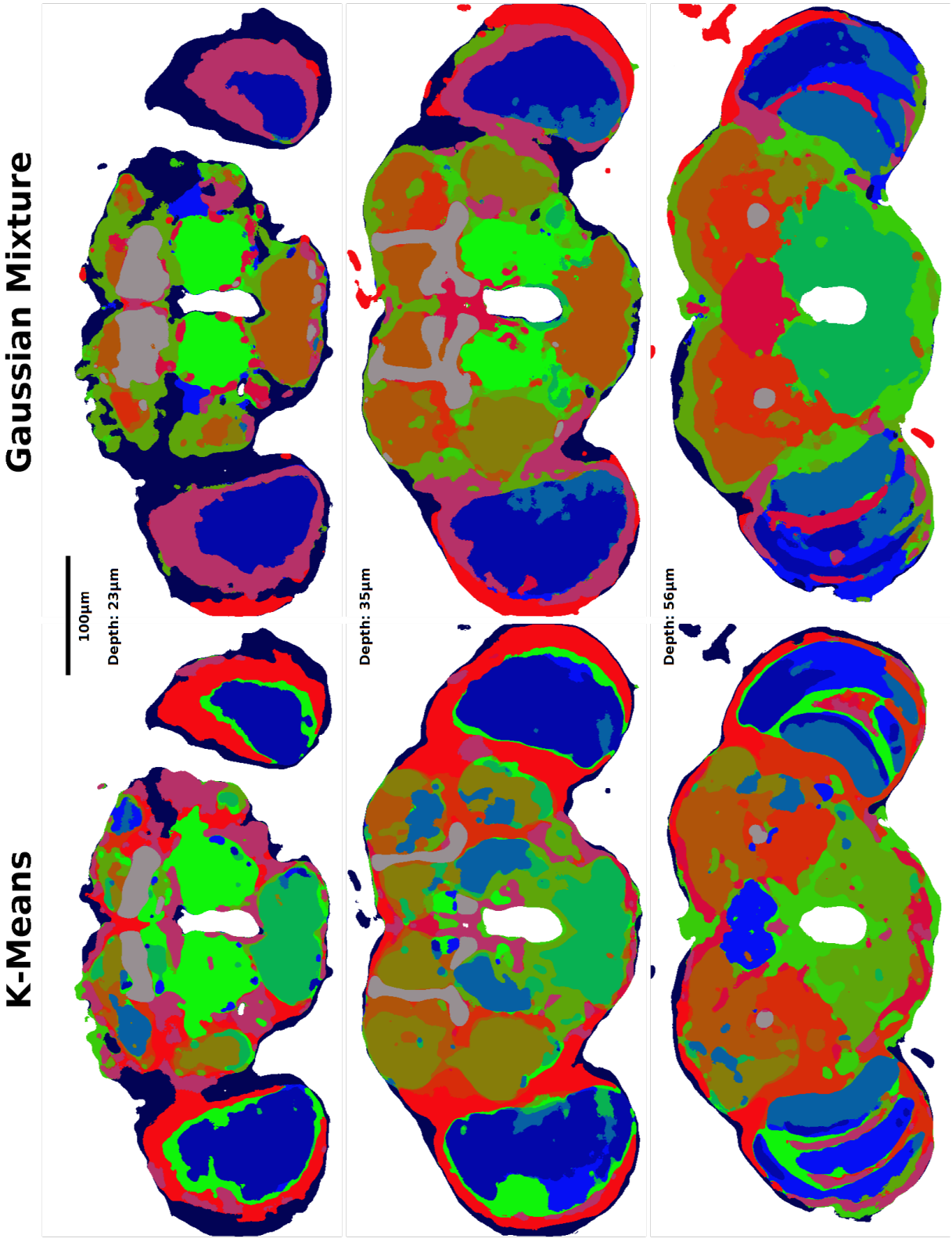


Figure 4.2: Natural segmentation of the *Drosophila* brain obtained by k-means and Gaussian mixture model clustering techniques. Results are shown for 15 clusters, based on selected start points throughout the brain. Similar results were obtained for different numbers of clusters and for random start points. Clustering was performed on 10 protein trap image stacks mapped to a male template brain selected from protein trap lines showing different expressions patterns.

are shown in Figure 4.2 (left). Seed points were located in the adult brain cortex, medulla, lobula, lobula plate, subesophageal ganglion, antennal lobe, calyx, gamma lobe, heel (of the mushroom body), ellipsoid body, fan-shaped body, protocerebral bridge, anterior ventral protocerebrum, superior lateral protocerebrum, and the lateral horn.

To assess the symmetry of the segmentation, a mirror image warp was calculated with CMTK to match the mirror image of the template brain nc82 channel to the non-mirrored version of the same brain. The resulting mirror-warp was applied to the K-means segmentation and each segment was compared with the mirror-warped version of itself. The Dice coefficient measure of set similarity was used to calculate segment agreement, where a score of 1 indicates perfect agreement and a score of 0 indicates no overlap.

$$s(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Despite using seed points from only one side of the brain the resulting segmentation was symmetrical, with an average Dice coefficient of 0.81, s.d. 0.07, averaged over all segments. Several distinct structures were segmented, including the medulla, lobula, mushroom bodies, antennal lobes and the central complex. In contrast to the FlyBase ontology, the noduli were not in the central complex segment, and the antennal lobes shared a segment with a thin layer in the optic lobes. Several other regions in the supraesophegeal ganglion were segmented, but these did not correspond well to the regions of the FlyBase anatomical ontology, as depicted on the Virtual Fly Brain website (Figure 1.9).

A Gaussian mixture model was also trained on the data with the same starting points, resulting in the segmentation in Figure 4.2 (right). Again the segmentation was symmetrical, with an average Dice coefficient of 0.83, s.d. 0.11, averaged over all segments. This method also segmented medulla, lobula, mushroom bodies, antennal lobes and central complex. In this case the noduli were included in the central complex and antennal lobes did not share any spatially separated segments. Segments corresponding to the clamp, and posterior slope, were also identifiable, despite no seed points being placed in these areas. The mushroom bodies were identified as one segment, despite being initialized with 3 seed points from different parts of the mushroom bodies. Compared to the k-means results, the Gaussian mixture model produced a segmentation more similar to the FlyBase anatomical ontology.

Randomly initialized segmentations, and segmentations based on additional pro-

Random Change	K-means	Gaussian Mixture Model
Seed Points	mean 0.79, s.d. 0.171	mean 0.579, s.d. 0.187
Brain Images	mean 0.642, s.d. 0.178	mean 0.642, s.d. 0.158

Table 4.2: Average segment consistency for segmentations repeated with random seed points, or random brain images. Average Dice coefficients are shown for each clustering method and each trial type. Each random trial was repeated 50 times.

tein trap images and clusters produced similar results. Slightly different groupings were obtained, some of which segmented structures such as the protocerebral bridge, layers of the fan-shaped body, and heel of the mushroom body from their neighboring regions. These segmentations are informative for how to segment regions of the *Drosophila* brain, based entirely on the similarities of their protein expression profiles.

To quantify the sensitivity of both K-means and Gaussian mixture models to changes in seed points and changes in input images, segmentation procedures were repeated 50 times with random starting points, and again 50 times with 10 brain images randomly selected from the 18 available expression patterns. Segmentation results for each trial were compared to the segmentations shown in Figure 4.2. The Dice similarity metric was measured between each segment in the original segmentation, $Orig_i$ $(i = 1 \ldots 15)$, and the new, random segmentation $Rand_j$ $(j = 1 \ldots 15)$. For each segment $Orig_i$, the "best match" segment in the new segmentation, $Rand_{best}$, was found so that the Dice coefficient between the two segments was maximized. The average $Rand_{best}$ score for all random trials was calculated to provide an indication of the stability or consistency of segment $Orig_i$.

Because of the uncertainty in brain registration, small misalignments can occur with an average of 4.49μm, s.d. 3.23μm, in neuropil regions (see section 2.4.2 for details). To allow for these small misalignments, a regional threshold of 5 pixels (2.44μm) was allowed for segmentation results based on randomly selected brain images. This threshold was not used for results based on randomly selected seed points.

Results averaged over all segments are summarized in Table 4.2, and for individual segments in Table 4.3. Overall, segmentations are not very consistent, however some segments are found consistently throughout the random trials, as shown in Table 4.3. For example, K-means segments 1 (outer adult brain cortex), 2 (medulla), 12 (inner adult brain cortex) and 15 (mushroom bodies) were more consistent throughout both types of trial, changing seed points and changing input brain images. Similarly, Gaus-

	K-means		Gaussian Mixture Model	
Segment	Seed Points	Brain Images	Seed Points	Brain Images
1	0.76	0.74	0.65	0.77
2	0.93	0.72	0.76	0.81
3	0.74	0.55	0.40	0.52
4	0.88	0.58	0.51	0.65
5	0.76	0.55	0.87	0.88
6	0.96	0.62	0.60	0.66
7	0.64	0.61	0.57	0.49
8	0.84	0.68	0.60	0.65
9	0.91	0.65	0.45	0.63
10	0.87	0.65	0.49	0.67
11	0.81	0.61	0.46	0.63
12	0.72	0.82	0.50	0.41
13	0.80	0.47	0.41	0.53
14	0.51	0.61	0.53	0.62
15	0.99	0.76	0.89	0.72

Table 4.3: Average segment consistency for individual segments after segmentations were repeated with random seed points, or random brain images. Average Dice coefficients are shown for each individual segment, grouped by clustering method and random trial type. Each random trial was repeated 50 times. Note that K-means and Gaussian Mixture Model segment numbers do not necessarily correspond.

sian mixture model segments 1 (adult brain cortex), 2 (medulla), 5 (subesophageal ganglion) and 15 (mushroom bodies) were more consistent than other segments.

Previous work on the mouse brain found that automatically derived segmentations had a high correspondence with the classical, manually labeled, anatomical reference atlas [9]. Because the segmentations derived here for *Drosophila* are based on only a small number of expression patterns they remain sensitive to initial conditions and to the expression patterns chosen. Despite this variability, several regions are consistent across many trials and different expression patterns, and these generally agree with the manually labeled FlyBase anatomical ontology. This suggests that using a larger set of protein expression patterns, it would be possible to produce a more robust protein-based segmentation, which would enable us to further understand the links between the level of molecular biology and our current anatomical understanding of the *Drosophila* brain.

4.3 Homology Networks

Protein trap expression information stored in the BrainTrap website is useful as a resource for studying protein expression in the brain. Additionally, the large amount of raw data on expression location is useful for further understanding genetic similarities between *Drosophila* and other organisms.

Genetic sequences allow us to compare genes between species. Anatomical protein and gene expression information allows further comparison of anatomical specificity and gene function between species. Furthermore, investigating differences in gene expression patterns between species informs us about how genes are preserved or modified and utilized in different ways throughout evolution. Studying the evolution of brain proteins in particular can inform us about the organization of the brain, and brain specialization [26].

Comparative analysis can also tell us how similar regions in the brain are to each other, or how they are similar between different species. Protein trap data in the BrainTrap database enables us to perform this analysis more accurately with *Drosophila*. In the mammalian brain, neuron cell bodies, where genes are expressed, are distributed throughout the different regions of the brain. This means that gene expression data, such as those recorded in the ABA, is sufficient to provide gross anatomical brain region profiles of genetic expression. In the *Drosophila* brain this task is more difficult because neuron cell bodies are all located in the outer layer of the brain, and send

neurites into the other areas of the brain. Protein localization is required to identify which proteins are specific to particular brain areas in *Drosophila*. While protein and gene expression have different properties, they both give anatomical expression profiles, and can be used to compare regions across species, at the anatomical level. In Section 4.3.1, I use the BrainTrap data and mammalian data, from the ABA [70], to identify brain regions that are homologous in gene / protein expression.

Comparative genomics databases contain a large amount of information on genes that are homologous between mouse and *Drosophila*. From the list of BrainTrap proteins, I identified a subset that have orthologues in mouse. Then from this subset I compared expression patterns between species. Homologous brain areas and general similarities between anatomical brain areas in mouse and *Drosophila* were identified, linked through 362 fly protein, and 544 mouse gene, expression patterns. Protein trap lines are selected at random with the piggyBac gene insertion method [43, 113] so this should give an unbiased coverage of the fruit fly genome.

Because we share a common ancestor, there are genetic similarities between mammals and insects. This allows us to study the operation of the same genes in different organisms. Anatomical expression data also gives us information with which we can measure aspects of evolution. Emes *et. al.* [26] previously investigated this with data from the ABA, where the evolution of brain gene "variability" was measured for ancient genes (with a shared homologue in yeast) and more recently evolved genes (without any homologue in yeast). In section 4.3.2 I replicate this analysis with fly brain protein expression data. Similar to the results obtained by Emes *et. al. [26]*, a trend towards more variability in recently evolved proteins was found.

4.3.1 Fly ~ Mouse Comparative Analysis

Using the gene expression information collected from the Allen Brain Atlas project it was possible to compare fruit fly neurobiology to mammalian neurobiology using homology networks to link the two anatomical expression pattern networks. Protein expression patterns in *Drosophila* and gene expression patterns in mouse measure different biological indicators, but both give a distinct expression pattern for each anatomical region of the respective brains.

To identify similarities between mouse brain regions and fly brain regions I correlated protein annotations from the BrainTrap database with homologus gene expression level data from the ABA. The Ensembl Compara database was used to identify homol-

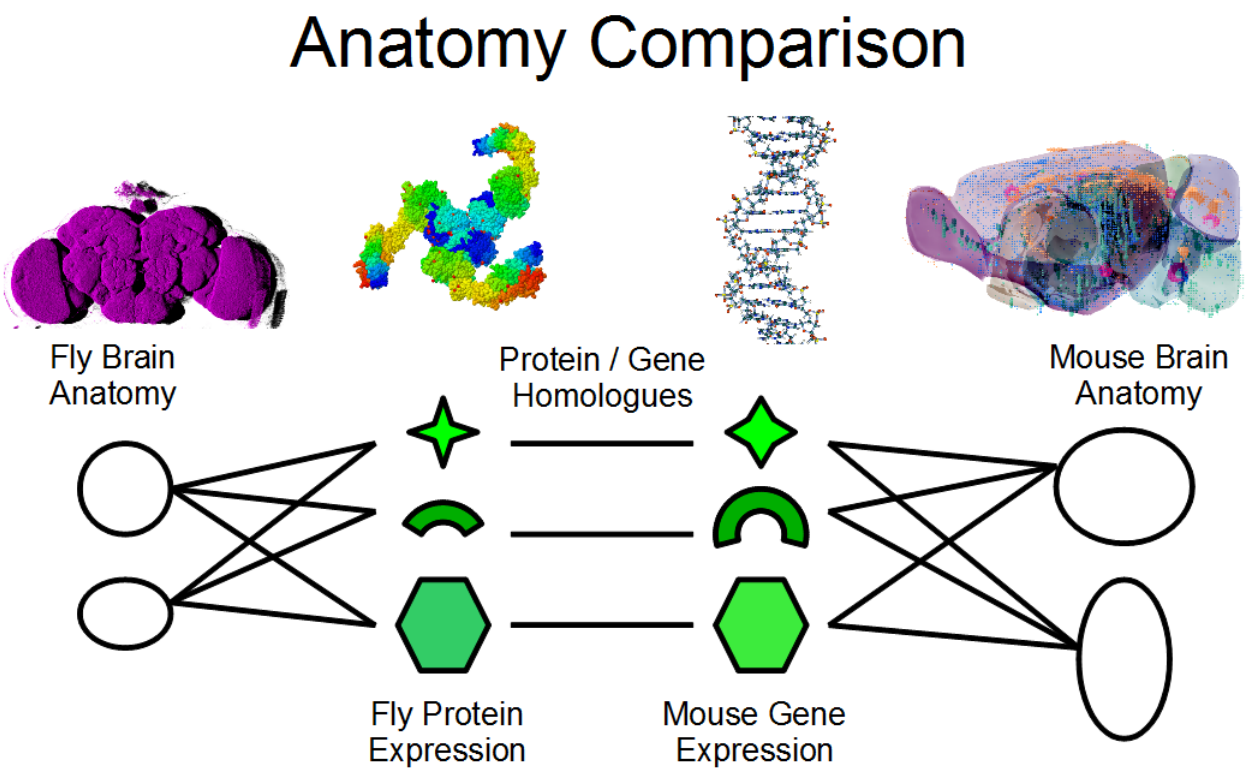


Figure 4.3: Fly Brain ~ Mouse Brain. This schematic summarizes the approach taken to compare brain regions between *Drosophila* and mouse brain anatomy. Fruit fly protein expression data from the BrainTrap database gives a protein expression profile for each brain region. Similarly, mouse gene expression information from the Allen Brain Atlas gives a gene expression profile for each brain region. Using the homology network to identify similar genes and proteins allows a comparison of brain region expression patterns across species.

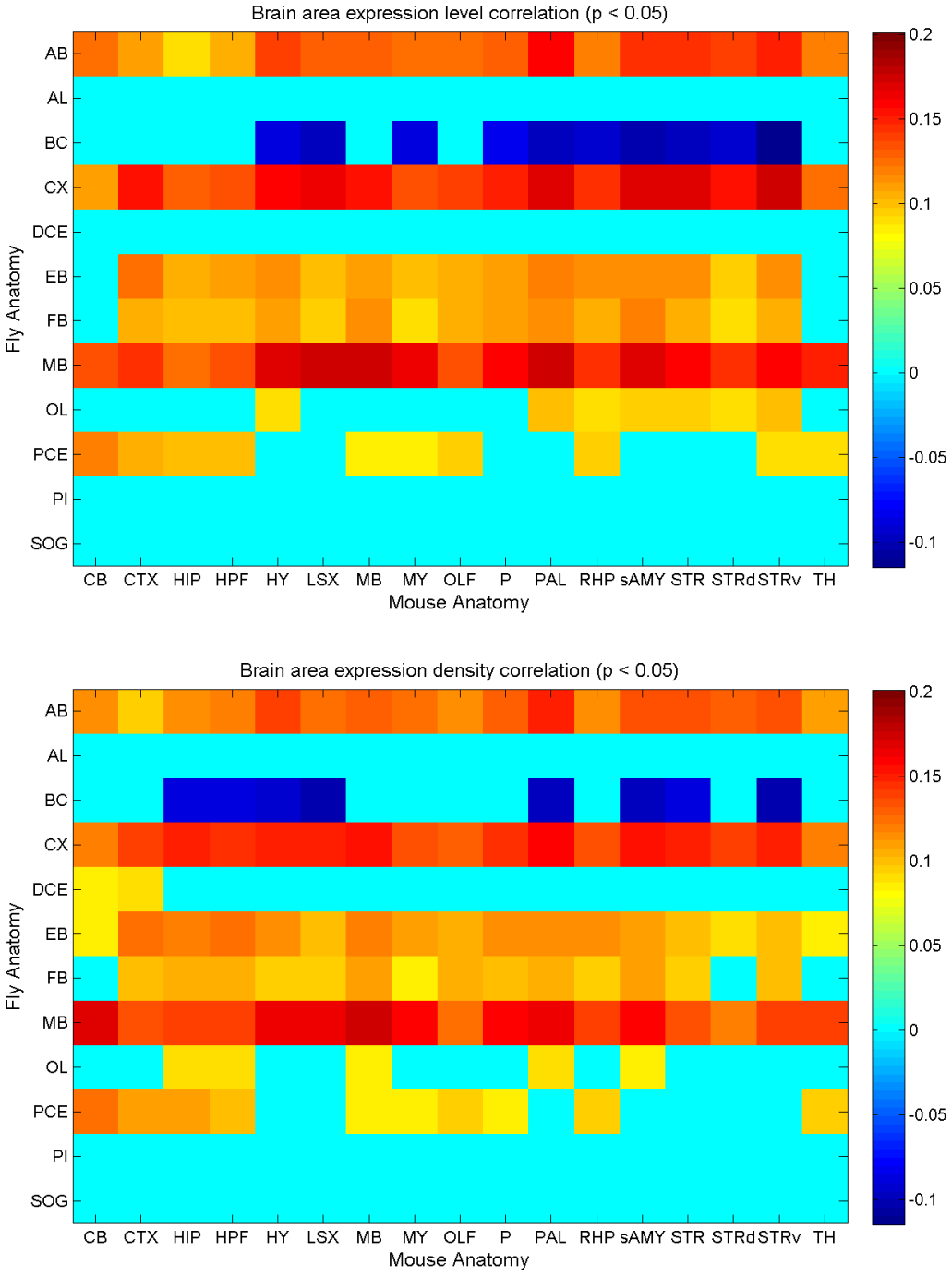


Figure 4.4: Correlations between fly brain protein expression, and homologous mouse gene expression patterns. Results are shown for BrainTrap annotation data correlated against mouse gene expression level (top) and density (bottom) from the Allen Brain Atlas (ABA). All correlations shown are significant to $p < 0.05$. All other correlations are shown as 0 correlation. Fly and mouse brain region abbreviations are explained in section 4.3.1.

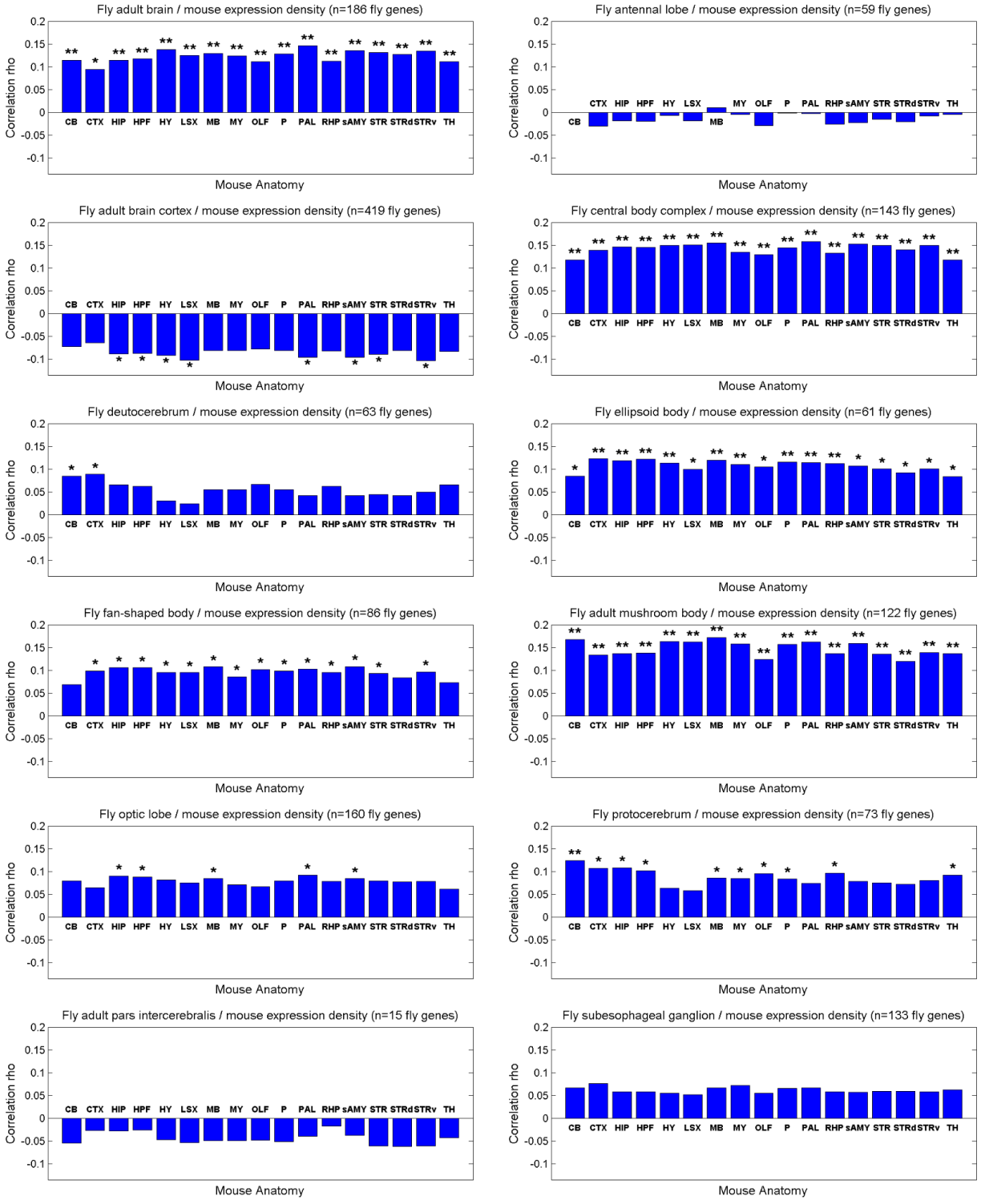


Figure 4.5: Correlations between fly brain protein expression, and homologous mouse gene expression patterns. Results are shown for BrainTrap annotation data correlated against mouse gene expression density from the Allen Brain Atlas (ABA). * indicates correlations significant to $p < 0.05$. ** indicates correlations significant to $p < 0.01$. Mouse brain region abbreviations are explained in section 4.3.1.

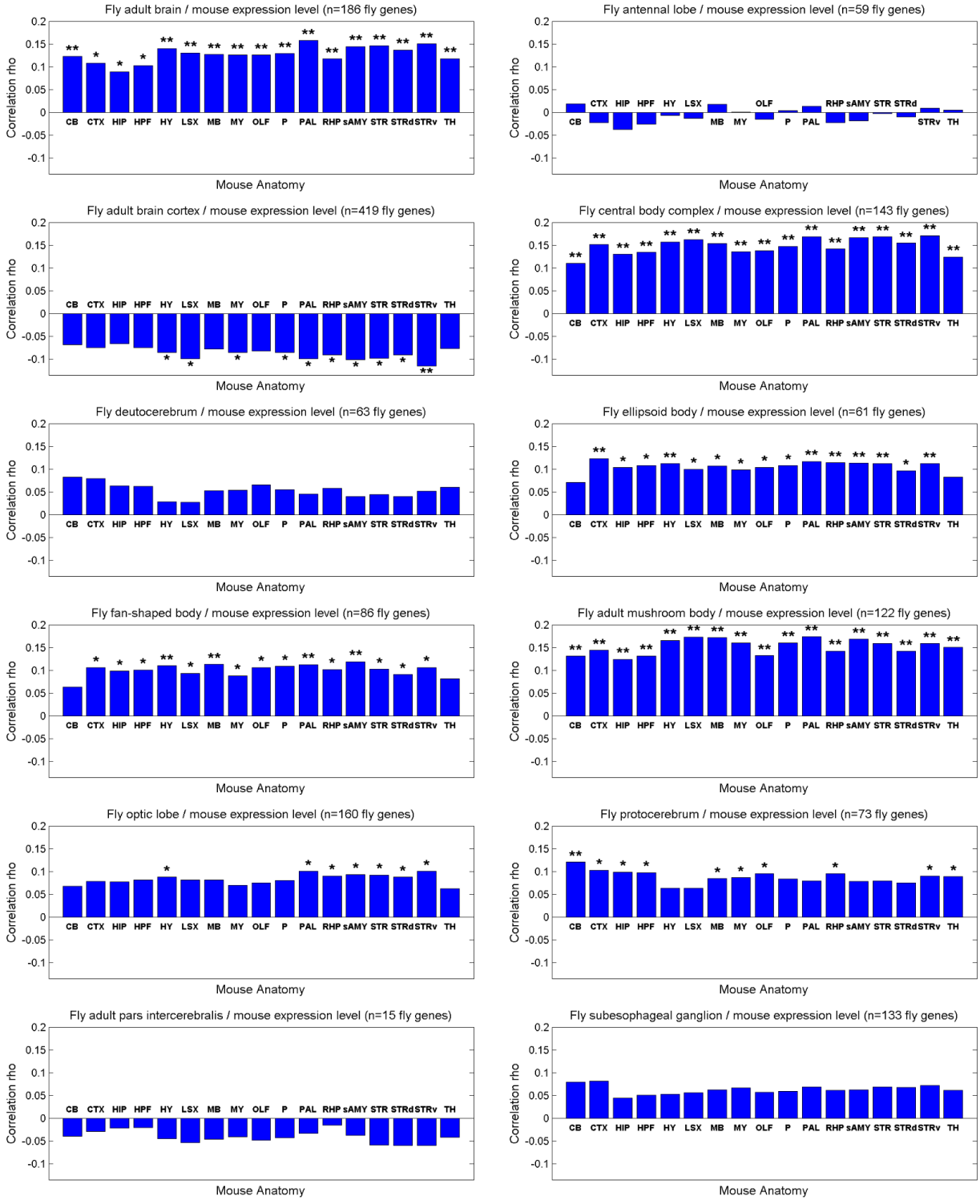


Figure 4.6: Correlations between fly brain protein expression, and homologous mouse gene expression patterns. Results are shown for BrainTrap annotation data correlated against mouse gene expression level from the Allen Brain Atlas (ABA). * indicates correlations significant to $p < 0.05$. ** indicates correlations significant to $p < 0.01$. Mouse brain region abbreviations are explained in section 4.3.1.

ogous genes in mouse from the subset of *Drosophila* genes in the BrainTrap Database. 362 fly genes were found, with 544 homologus mouse genes. These 544 gene homologue pairs were used as a basis for anatomical comparison across species, as depicted diagrammatically in Figure 4.3. 12 *Drosophila* anatomical areas commonly annotated in the BrainTrap data were selected: adult brain (AB), antennal lobe (AL), adult brain cortex (BC), central body complex (CX), deutocerebrum (DCE), ellipsoid body (EB), fan-shaped body (FB), adult mushroom body (MB), optic lobe (OL), protocerebrum (PCE), adult pars intercerebralis (PI), and subesophageal ganglion (SOG). For each of these areas proteins were assigned 1 (expression present) or 0 (no expression).

Similarly for the mouse brain, 17 major areas from the Allen Brain Atlas anatomical regions were used: cerebellum (CB), cerebral cortex (CTX), hippocampal region (HIP), hippocampal formation (HPF), hypothalamus (HY), lateral septal complex (LSX), midbrain (MB), medulla (MY), olfactory bulb (OLF), pons (P), pallidum (PAL), retrohippocampal region (RHP), striatum-like amygdalar nuclei (sAMY), striatum (STR), striatum dorsal region (STRd), striatum ventral region (STRv), and thalamus (TH). For each of these regions the average expression level and density were calculated from the ABA data. It is not clear which measure (expression level or expression density) is more appropriate to use for this analysis, so results for both measurements are shown.

For each anatomical region pair, Pearson's correlation coefficient (rho) and significance test p-values, were calculated. Significant ($p < 0.05$) correlations are summarized in Figure 4.4, and individual results for each fly brain region are shown in Figures 4.5 (expression density) and Figure 4.6 (expression level).

Several pairs of brain regions show significant correlation. The adult brain (a tag used to identify proteins expression throughout the brain) showed significant correlation with all mouse brain areas, as did the central complex and the mushroom bodies. This general correlation is also seen for the ellipsoid body and fan-shaped body (parts of the central complex). The adult protocerebrum shows correlation with mouse brain areas such as the cerebellum, cerebral cortex and hippocampus. Compared to regions such as the antennal lobe or subesophegeal ganglion (where there is no significant correlation) these *Drosophila* brain regions have more in common with parts of the mouse brain.

The adult brain cortex shows negative correlation with several brain areas including the hypothalamus, lateral septal complex, and the striatum. As discussed in section 3.1.2, proteins that localize to to the adult brain cortex in *Drosophila* are less likely to

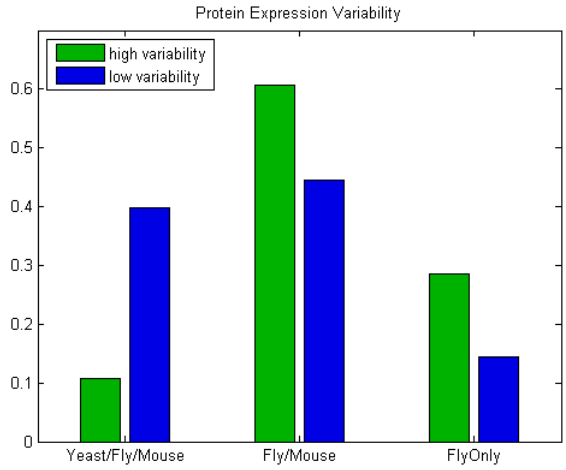


Figure 4.7: Evolutionary brain protein expression variability. Ancient genes, shared with yeast, are more likely to have low variability throughout the brain. More recently evolved genes, observed in fly and mouse or fly only, are more likely to have a high variability within the brain.

be involved in important brain function. This suggests a general negative correlation showing that proteins less important for brain function in *Drosophila* are also less important for brain function in mouse. Combined with the positive correlations seen for other neuropil regions this is encouraging for *Drosophila* brain protein research.

4.3.2 Variability Yeast / Fly / Mouse

Previous research has compared gene expression variability between yeast and mouse, showing the development of brain signaling proteins from yeast to mammal [26]. Using data from the BrainTrap screen and homology networks from Ensembl Compara, I analyzed data from *Drosophila* in the same way, linking genes both to yeast and to mouse.

Firstly I defined variability in terms of the anatomical tags in the BrainTrap database. Genes that were tagged with expression throughout the adult brain, or in the adult brain cortex were considered low variability. Genes tagged with expression in other areas were considered specific.

Next, using homology data from the Ensembl Compara database, each protein in the BrainTrap database was classified as either having homologues in yeast and mouse (Yeast/Fly/Mouse), homologues in mouse only (Fly/Mouse), or having no known homologues (FlyOnly). The number of high or low variability genes were counted for

each category and results are shown in Figure 4.7.

Generally expressed genes (with low variability) are more likely to have homologues in yeast. Specific genes are more likely to have homologues in mouse or no homologues at all (fly only). These results, based on fruit fly data, reflect the results from [26], based on mouse data, and demonstrate the action of evolution. More ancient genes which are shared with yeast are more likely to have low variability throughout the brain, and newly evolved genes tend to have a high variability within the brain, giving a detailed, more specific role for the genes that evolved later.

4.4 Connectivity Networks

Connectivity data is very difficult to collect (see Chapter 5 on tracing neurons). Recently there has been much progress in this field, and many neurons have been traced through various areas using light and electron microscopy techniques. In *Drosophila* there is a bias in the available connectivity data towards well studied areas. Areas such the optic lobes and the adult central complex have received much attention and many connections have been recorded, but other areas such as the unstructured neuropils are less well mapped out (Figure 4.8). Many parts of the unstructured neuropils in the protocerebrum are responsible for different behavior between sexes [86, 49, 60, 12]. With data collected and available online [119] it is now possible to analyze connectivity data in combination with anatomical expression to add information to the connectivity network. In particular, synaptic and neurotransmitter / signaling molecule distribution can tell us more about the properties of connections in each brain region. Unfortunately obtaining information on many synaptic connections at once requires EM connectivity maps, which are much more difficult to collect and annotate.

Many experiments on light level anatomical tract data are available in published literature, and some of this data has been combined into a connectivity database [119]. Figure 4.8 shows the limitations of this data; regions are connected by observable tracts of neurons, but in most cases direction of connectivity is not known. At present more connectivity data is required for some regions, and light level connectivity of this kind is not able to measure neuron-to-neuron connectivity for a large number of cells, so analysis of this type of data is limited. In Section 4.4.1, I demonstrate one way this data could be combined in future, by performing cross-species analysis on a small amount of connectivity data. In the future, combining protein expression data with connectivity data such as this could be very informative about the role of proteins in

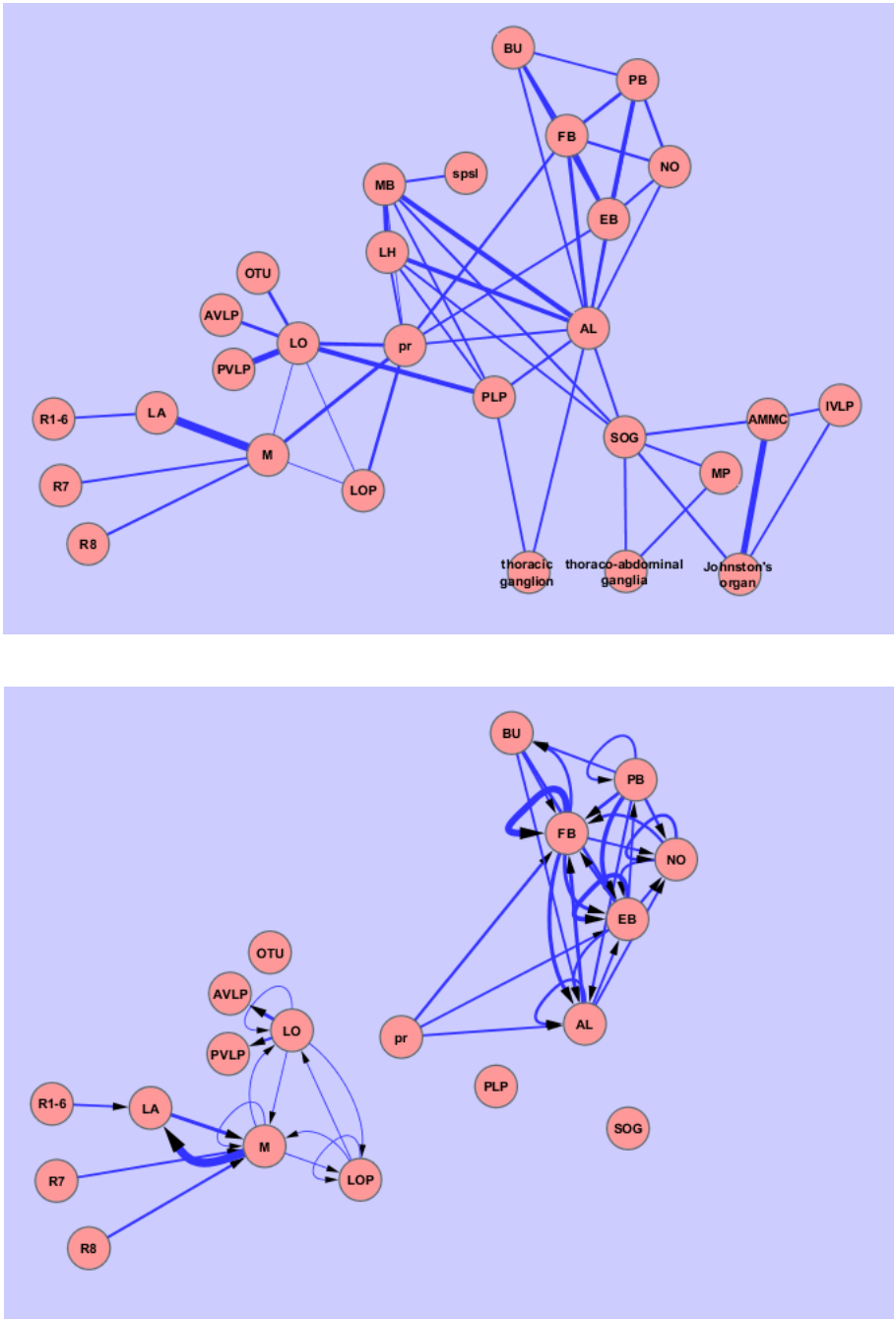


Figure 4.8: Undirected (top) and directed (bottom) region connectivity in the *Drosophila* brain, from data collated in the Flybrain Neuron Database [119]. Line thickness represents the number of known connections from one region to another. Compared to undirected connections, the directed connections are not well studied, and are biased towards areas such as the optic lobes and the central complex.

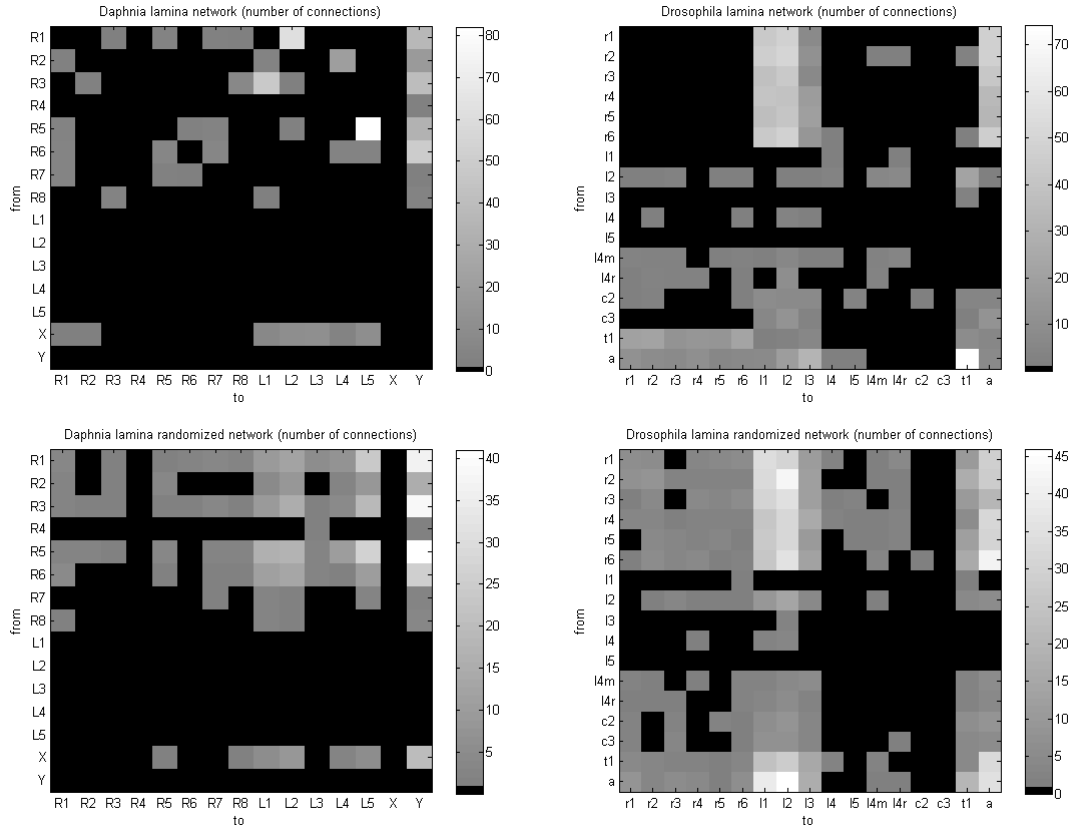


Figure 4.9: Top: Connectivity networks of *Drosophila [84]* and *Daphnia [79]* lamina. Brightness indicates the number of synaptic connections between each neuron, according to the keys on the right. Bottom: Randomized connectivity networks of *Drosophila* and *Daphnia* lamina. Brightness indicates the number of synaptic connections between each neuron, according to the keys on the right. Despite preserving the same number of inward and outward connections for each neuron the randomized networks are visually very different from the original networks.

wiring up the brain.

4.4.1 Enhanced Motifs

Motif analysis has emerged as a popular way to compare the basic building blocks for different networks [87]. A diverse range of network types, from computer circuitry and social networks to protein interaction and neural connectivity networks. Motif analysis is a useful tool, but it is limited in the amount of insight it can provide. It is unclear what the motif results tell us about different connectivity strategies. Here I present a new method for adding node property information to motif analysis to

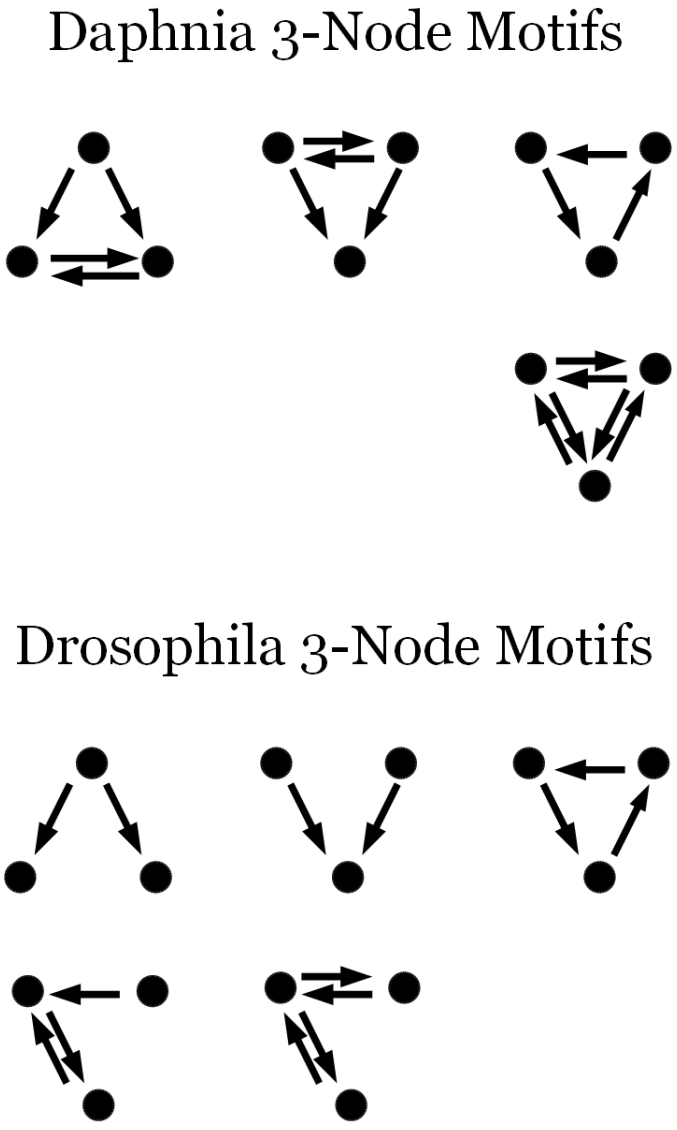


Figure 4.10: Three-node network motifs over represented in reconstructions of *Drosophila* and *Daphnia* lamina.

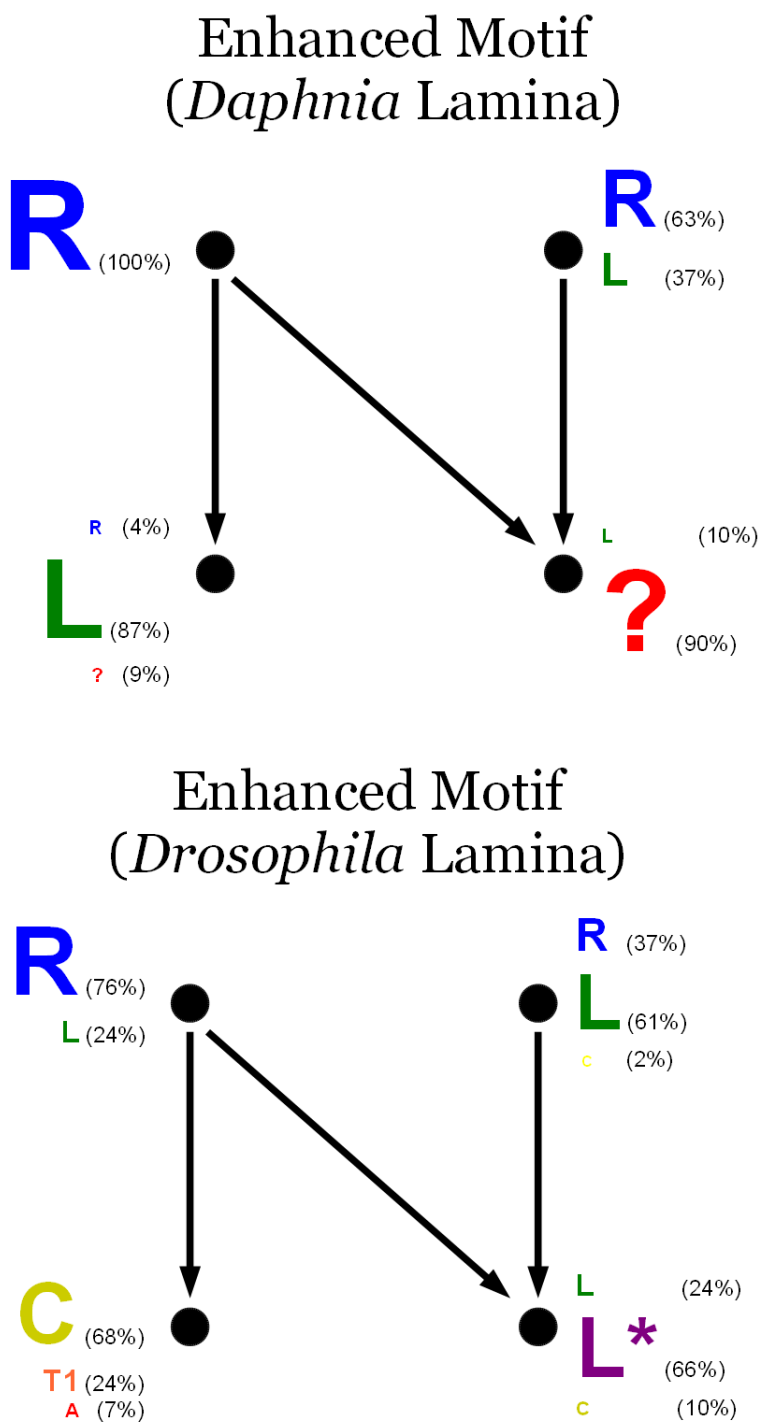


Figure 4.11: Enhanced motifs for *Drosophila* and *Daphnia* lamina show more information. Letters denote neuron type R: receptor cells, L: lamina cells, L*: Lamina cells from a neighboring cartridge, C: central cells, A: Amacrine cells, T1: T1 transitional cells, ?: unknown cell type or untraced cell. These N-shaped motifs are found to be over-represented compared to random networks in both Drosophila and Daphnia networks. Additionally, different cell types show a preference for different locations in the motif.

provide further insight to the similarities or differences between, in this case, neural connectivity strategies.

Motif analysis is a way of analyzing information from large networks of data. Occurrences of small patterns of connectivity between nodes in the network can be compared (Figure 4.10). Motif analysis has previously been used to compare many types of network and has found common motif patterns in computer circuitry, gene networks, food chains and neural networks [87]. The technique can also be used to analyze neural connectivity collected from EM images, such as the *C.elegans* connectivity network, which was analyzed in [108]. Node type information is discarded for this type of analysis, and in the case of neural connectivity, inhibitory and excitatory neurons are treated in the same way; so although motifs are powerful tools for network comparison they ignore much important information included in the network.

The following method was developed to enhance the motif diagrams with node type information. This provides more information about the types of node that contribute to each position in the motif. It also provides a more powerful way of comparing networks of the same type, revealing additional similarities or differences.

Following the notation used in [108], let the number of N-node patterns in the i-th motif class in the actual network A be defined as $c_{N,i}(A)$. Also define the nth node in pattern m of this motif class as having type $T_{n,m,N,i}$. Then after calculating significantly overrepresented motifs [87, 108], for a significant motif of class i, measure for node n in the N-node motif, the proportion of nodes of type t contributing to that motif at that location:

$$p_{i,n,t} = \frac{\sum_{m=1}^{c_{N,i}(A)} \begin{bmatrix} 1 : T_{n,m,N,i} = t \\ 0 : otherwise \end{bmatrix}}{c_{N,j}(A)} \tag{4.1}$$

Note that a single node can contribute to a motif class multiple times in the same location within the motif, so it is possible for some nodes to be counted more than once in this calculation. Also, when two or more nodes in a motif can be swapped without changing the motif, the sum should be repeated for all such nodes, n_1 and n_2 for example, and the total normalized by $x * c_{N,i}(A)$, where x is the number of interchangeable nodes.

Repeating the calculation for each node type we can construct an enhanced motif representation as shown in Figure (4.11), and this can be repeated for all significantly overrepresented motifs in the actual network. This procedure can also be repeated for

motifs that are not significantly over represented in the actual network, however this is not likely to be as insightful.

To demonstrate this technique I compared the connectivity networks of the lamina of *Drosophila* [84] and *Daphnia magna* [79] (Figure 4.9, Top). Following the methods described in [108] random networks were generated with the same number of unidirectional and bidirectional connections for each node (Figure 4.9, Bottom). Motif frequencies in the real network and the random networks were compared and motifs overrepresented in the real network were identified (Figure 4.10).

Similar motifs overrepresented in two networks of neurons does not necessarily mean the connectivity strategies are also similar; information on types of neurons contributing to each motif are not considered. One way to include extra node type information in the motif diagrams is to show the relative contributions of each type to each node in the motif (Figure 4.11). This allows further, more detailed comparisons of connectivity strategies between species.

In future, this approach could be applied to neural connectivity data combined with protein or gene expression data, neuron type data (as shown in Figure 4.11), inhibitory / excitatory data, or neurotransmitter type data. Neuronal connectivity data at this level of detail is rare and, where it does exist, networks are very small, with the exception of the *C.elegans* neural reconstruction [136]. Therefore, the application of this technique to neural data is currently limited. In Chapter 5, I develop techniques to assist researchers in the semi-automatic annotation of EM data required for this type of reconstruction.

4.5 Summary

In this chapter I developed techniques for linking protein localization information collected from the brain of *Drosophila* with the levels of anatomy and connectivity. This approach was informative about possible protein-protein interactions, and about the similarities of *Drosophila* brain regions at the molecular level. Taking an evolutionary interpretation of the results it was possible to link regions of the fly brain to regions of the mouse brain, based on similar protein and gene expression profiles. It was also possible to reveal in greater detail the way that evolution has shaped the variability of protein expression in the brain.

Neuron type data analyzed in conjunction with neural connectivity also has the potential to uncover further information about the wiring specificity present in brain

networks, as demonstrated by the enhanced motifs discussed in Section 4.4.1. The amount of connectivity information is currently very limited and this technique will be more powerful when further data is available for enhanced motif analysis.

Research into functional connectivity, and connectivity imaging techniques is receiving renewed interest in the scientific community [10, 74, 19], so we can expect more connectivity data to be available in the future. In the next chapter I develop methods for improving the automatic annotation of this connectivity data, inspired by the properties of biological vision.

Chapter 5

Image Analysis and Neuron Tracing

All brains, from insects to mammals, consist of an intricate, densely interwoven 3D structure in which neurons connect to each other in a specific way to produce the diverse range of behavior we observe in nature. The pattern of connectivity in any brain is fundamental to the function of the neurons and the production of behavior. Because of the difficult microscopy techniques required to measure the full synaptic level connectivity between neurons, neuroscience has not yet been able to measure the full connectivity of any organism other than *C. elegans* [136]. Such high resolution connectivity data provides a basis for modelling and improving our understanding of the workings of these nervous systems. This type of data can form the basis of well constrained computational models, as demonstrated with models of the crustacean stomatogastric ganglion [118, 125, 124], and *C. elegans* nervous system [125, 124, 55].

5.1 Motivation

Recently, a connectivity research renaissance has reawakened investigation of the connectivity patterns in mammalian and insect brains [10, 74, 19]. Several labs have recently used connectivity data to study functional connectivity, relating the structural connectivity of neurons in the brain to the functions they perform [8, 11]. Both light and EM images are employed by these experiments to determine the function of neurons and measure the dense connectivity within the brain. These papers show that the structural and connectivity strategies employed by the brain are measurable and important for the functions performed by neurons. Other labs have used powerful genetic techniques to image neurons at the light level and trace the path of neurons at the neuromuscular junction [73, 77], in mouse cortex [76], and in the fly brain [36, 37].

Because of the stereotypical layout of neurons in the fruit fly brain it is possible to repeat these experiments in different brains and combine the results into a reference brain where many neural projections from different experiments can be examined at once.

Currently this research requires many hours of manual annotation to trace the locations and connectivity of neurons. To trace neurons imaged with light-based microscopy a researcher must manually trace the branching tree-like structure of single neurons, or in some cases several differently coloured neurons. For neurons imaged with electron microscopy, researchers, and sometimes large teams of annotators, must manually follow the paths of each neuron through the volume and record the path taken and the synaptic connections made in order to determine the connectivity inside a volume of tissue. This presents two challenges to computer science. Firstly, and ideally, we need to provide algorithms capable of automatically annotating these images of neural tissue. Current algorithms for such image annotation are far from perfect [74, 19]. Therefore the second challenge is to provide effective interfaces to allow human annotators to correct the partial annotations provided by the automated methods. I have contributed towards solving these challenges by investigating a novel method for tracing membrane in EM images and improving alignment between consecutive ssTEM images, and developing an example of a web interface for correcting automatically generated ridge detection results.

5.2 EM Level Tracing

In order to assist in the task of manual annotation I investigated methods for automatically detecting sections of membrane and aligning consecutive ssTEM images based on these automatic detections. Some of this work also appears in the published paper "Biologically inspired EM image alignment and neural reconstruction" [62], included as Appendix 2. Here I will discuss the algorithms used in more detail, and provide examples of possible extensions or alterations. Section 5.3 includes details of the web interface for manual correction of the automatic membrane detection, and Section 5.4 describes an extension of the tracing algorithm to work with 3D receptive fields implemented on a GPU to trace neurons in near-isotropic EM images.

Serial-section transmission electron microscopy (ssTEM) can produce reconstructions of neuronal morphology at very high resolution, including synaptic organelles and the contacts between neurons that constitute circuits. Alignment and reconstruc-

tion of ssTEM images is currently performed manually or semi-automatically, with the aid of computer software, to generate a 3D model of the imaged neuron and, with other such neurons, of the synaptic circuits to which that neuron contributes [136, 84, 126]. In some cases approximate alignment can be achieved automatically but, even so, high quality circuit reconstructions still require many hours of manual tracing and annotation, and are highly dependent upon the interpretative skill of the human observer and the complexity of neuronal arborizations being reconstructed [10, 121, 25]. For example, neurites are simple and have been completely reconstructed by manual means in the nematode *C. elegans* [136], but are complex, highly branched and reconstructed only with great difficulty or incompletely in the fruit fly [126].

Existing methods of image alignment usually rely on a control point selection method. Semi-automatic alignment can be carried out by identifying control points manually and aligning these by means of a particular algorithm [66, 28]. Automatic alignment can be carried out with control point detection algorithms [3, 115], and works best when image quality is consistent throughout the data, but performance can be degraded when artifacts such as gaps, noise, or differing levels of brightness and/or contrast are present in the images being aligned. Such artifacts are unfortunately fairly common in ssTEM images because of the complicated preparative procedures and increase as the series of images becomes longer.

ssTEM images are a series of high-resolution two-dimensional images that are not aligned by default. Some algorithms exist to align these images, but alignment is only approximate. I have tried a receptive field based approach for identifying and tracing membrane in a set of serial section EM images generated by Zhiyuan Lu and Ian Meinertzhagen, taken from the calyx of the mushroom body in a *Drosophila* brain [71].

We explored a novel approach based on receptive fields to identify the likely locations for cell membranes, and a ridge detection approach to identify lines within the receptive field responses. We then used a shortest-path approach to close the edges detected in the image. Lines of cell membranes are aligned in 3D to improve the image alignment and generate partial 3D reconstructions. Detected membrane points are further analyzed to identify likely organelle and synapse locations within the ssTEM images.

Receptive fields are a well-studied feature of many sensory interneurons, especially in visual systems, and define a region within which the neuron responds to a particular stimulus, such as a line segment at a particular orientation [44, 92, 31, 4]. Compu-

tational models of the brain also use receptive fields to further understand the visual system [93, 94, 13]. In the system presented here, receptive fields are learnt from examples of image patches taken from ssTEM training data using supervised learning techniques. The resulting Gabor-like receptive fields are applied to ssTEM data images to automatically annotate neuronal membranes, synaptic connections, and organelles such as mitochondria. These receptive fields are applied to ssTEM images with local inhibition, and excitation for congruent receptive fields, similar to the lateral inhibition and excitation observed in biological vision systems [38, 4, 98, 95]. Objects recognized by the system are then used to improve the alignment of consecutive images and produce partial 3D reconstructions as a starting point for manual annotation. Using this biologically inspired approach to analyse and understand biological images has the potential for further improvements in semi-automatic segmentation by applying additional properties of biological vision.

Membrane receptive field responses were used as a basis for a ridge detection algorithm as an alternative to the widely-used watershed algorithm. Results from ridge detection were used to improve alignment between slices with the use of a new dynamic programming procedure for aligning sequences. A variation of the Smith-Waterman DNA sequence alignment procedure [122], with a cost metric based on the euclidean distance and angular subtense, was developed for this purpose.

Results are still far from perfect, but objects recognized by the system can be used to improve the alignment of consecutive images and produce partial 3D reconstructions as a starting point for further manual annotation. A simple web annotation system has also been developed to manually correct partial reconstructions and output full 3D reconstructions. Based on manual corrections, membrane detection performance was assessed, as shown in Figure 5.9. It was possible to detect 95.5% of membrane present in the data with only a 13.9% false positive rate, or achieve a Rand index of 79.1% (Table 5.1).

5.2.1 Receptive Fields

Over 500 examples of membrane, synapses and mitochondria from serial 50nm thick sections of the mushroom body calycal neuropiles of *Drosophila* were manually annotated (Figure 5.1). The images were generated by Zhiyuan Lu and Ian Meinertzhagen, Dalhousie University, see [71] for details.

Training images were first normalized to have a range between -1 and 1. Then line

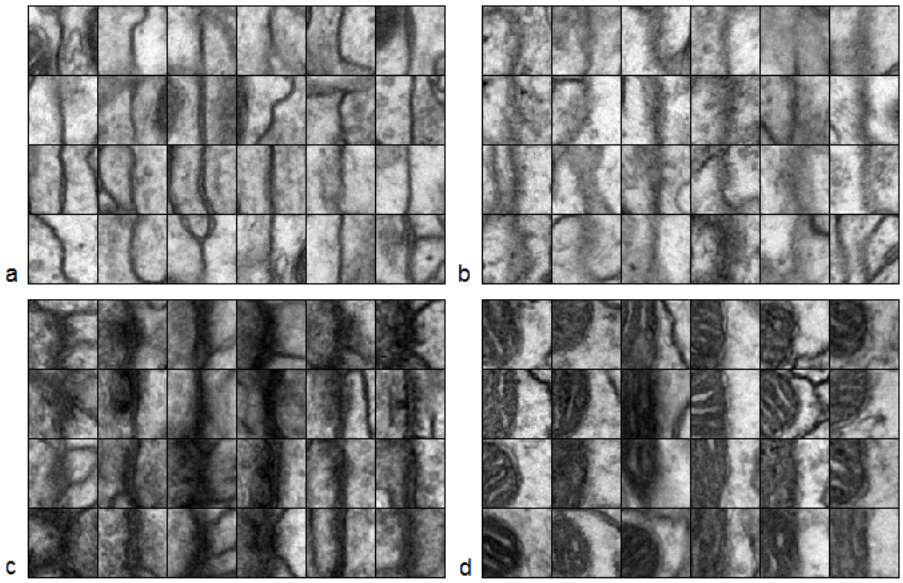


Figure 5.1: Examples of manually annotated training image patches from 50nm thick EM images of the calyx of the *Drosophila* mushroom body [71]. Four categories have been annotated: sharp membranes that pass vertically through the section thickness (a), blurred images of obliquely sectioned membrane (b), synaptic profiles (c) and mitochondria (d). Each image patch is 227x227nm.

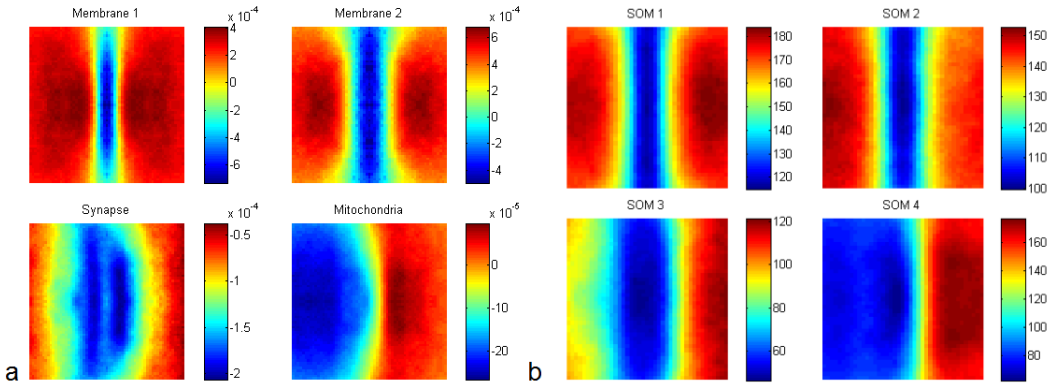


Figure 5.2: Weight matrices obtained after training on over 500 manually annotated image patches, such as those in Figure 5.1, and 400 randomly selected image patches from the same data. Supervised learning using a pattern recognition neural network produced results shown in (a). Unsupervised learning using SOM produced similar results (b). Weight matrices from (a) were used as a basis for the membrane detection algorithm. Pixel colours represent weights from input pixels, as shown in the right hand keys.

detection Gabor filters at different orientations were used to detect the best orientation for each image automatically. Images were rotated so that a vertical Gabor patch produced the largest response. In the case of images of mitochondria and synapses, rotation was performed to orientate the darkest half of the image to the left. Examples of resulting image patches are shown in Figure 5.1.

In the next step, a neural network was trained on the manually annotated image patches, randomly split into training and test groups, and a selection of 400 random images from the same data. A standard feed forward back propagation neural network was used, with a single input per pixel and as many outputs as target classes. All weights and biases in the network were initialized to zero and mean squared error was used as the error function. Membrane receptive fields were trained by specifying just one target class and training on membrane (or oblique membrane) from the training images with random images provided as negative examples. Synapse and mitochondria receptive fields were trained at the same time by specifying two target classes and training on positive examples from the manually annotated data, with membrane and random images used as negative examples. Training continued until classification performance on the test images stopped improving, typically after 10-20 iterations. Resulting weight matrices for each class are shown in Figure 5.2a. Note that a self-organizing map (SOM) unsupervised learning system also produced similar results and was able to identify membrane, synapse and mitochondria classes, as shown in Figure 5.2b.

5.2.2 Segmentation

Generic filtering can be improved by using machine learning to customize filters, as described in Section 5.2.1. The watershed algorithm is commonly used to segment EM images, but it usually results in over-segmentation [19]. Here we explore a ridge detection alternative to the watershed algorithm, which incorporates elements inspired by lateral inhibition and excitation observed in biological vision systems.

Space-filling three-dimensional reconstructions require us to identify the surfaces of neurons. These are necessary as the first step in identifying sites of contact between reconstructions of neighbouring neurons. The two main steps usually performed for EM image segmentation are filtering the image with an appropriate set of filters, followed by processing of the filter responses to obtain a segmentation. Usually generic filters, such as Gaussian smoothed Hessian (GSH), are used because they are fast to

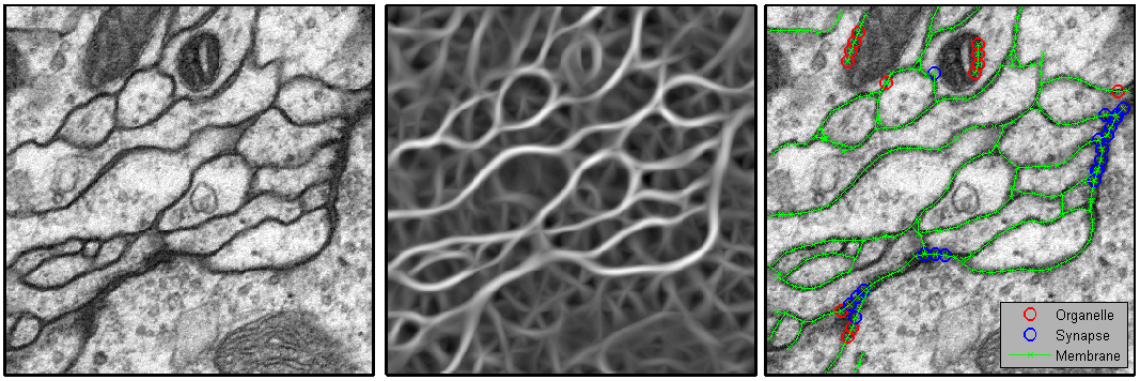


Figure 5.3: Membrane detection steps. Left: original EM image. Middle: maximum filter responses (combined over all angles). Right: results from automatic membrane detection and mitochondria / synapse classification results. Green lines indicate automatically detected membrane, red circles highlight automatically detected membrane that is classified as organelle (and therefore possibly false positive), and blue circles highlight automatically detected membrane that is classified as synapse. Each image is 1.4x1.4µm.

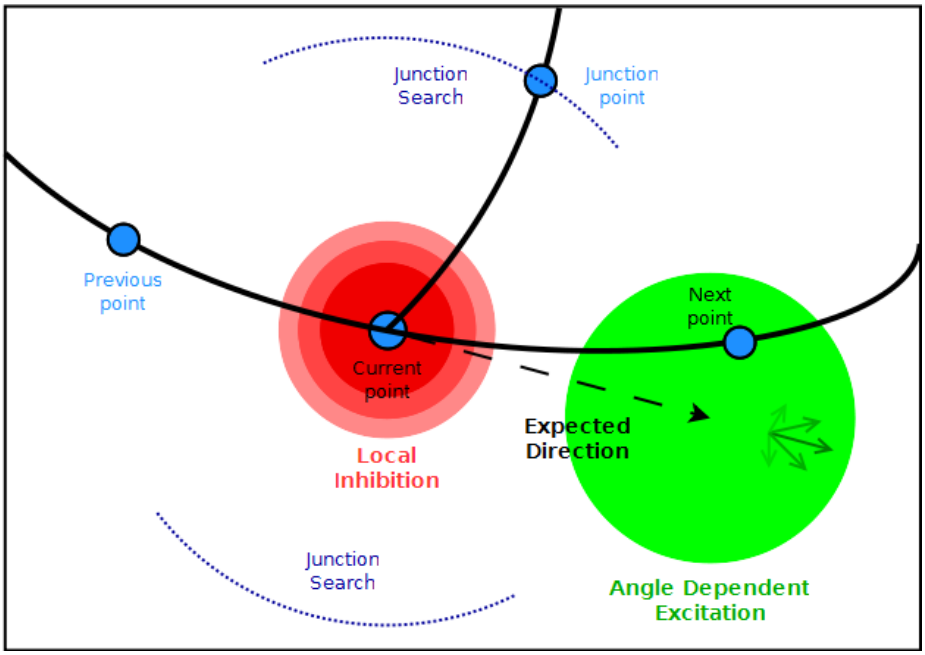


Figure 5.4: Ridge detection schematic for the step-wise ridge detection, based on receptive field responses. The area around the “Current point” is inhibited, and angle-dependent excitation is applied in the expected direction. A search is performed in this area for a local maxima, to find the “Next point”. Junction points are found away from the expected forwards and backwards directions.

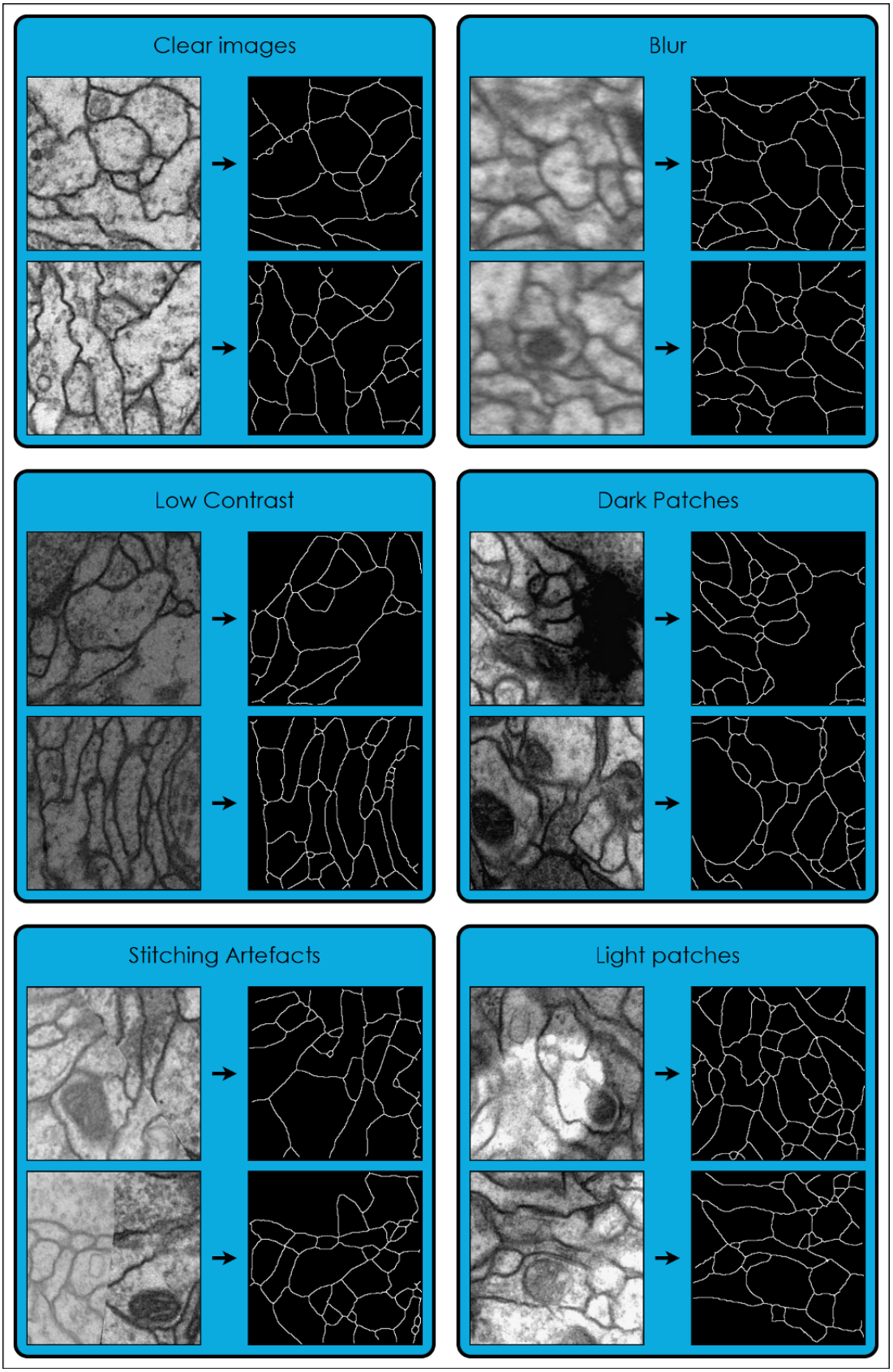


Figure 5.5: Ridge detection results in areas of noise commonly found in ssTEM images. This method is robust to many types of noise found in the images. Receptive fields are relatively invariant to changes in contrast and brightness, and edge closure can join gaps across small areas of noise.

calculate, and for segmentation the watershed algorithm is the standard approach [19].

As the basis to detect membranes, we used the membrane weight matrices from the neural network training to build a membrane detection tool. Weights were multiplied with a Gaussian probability density having a standard deviation chosen so that a constant background input produced a response of 0. The resulting Gabor-like patches were rotated to create a filter bank of membrane receptive fields. A total of 36 orientations was sufficient to produce reasonable accuracy to detect membranes (Table 5.1). Filters were convolved over large image patches from the same data to produce filter response images as shown in Figure 5.3 (middle panel).

Next, a ridge detection method was used instead of the watershed algorithm. The approach is depicted schematically in figure 5.4. Start points are initially identified as local maxima in the receptive field responses, above a chosen threshold. At each point, local inhibition is applied to receptive field responses and lateral excitation is applied at the expected next point. Excitation is angle dependent, so that continuous lines are more likely to be detected. Modifying the scores for nearby receptive field responses in this way can also be thought of as an approximation to the lateral inhibition and excitation observed in biological visual systems [38, 98, 97, 45, 99]. The next point is then found by searching for the resulting maximum in an area near the expected next point. A search is also conducted for junction points, which can be located in two arcs away from expected forwards and backwards directions for the current line. This search was repeated in a step-wise manner to propagate the current line of membrane progressively until a lower bound at some limit was reached. Neighbouring areas of high response with different orientations were marked as potential junction points and investigated in the same manner. An example of the results from this membrane detection process is shown in Figure 5.3, right panel.

This ridge detection method can join membrane over small distances of difficult to identify membrane, even in the presence of noise often found in ssTEM images (Figure 5.5). With a small amount of gold standard segmentation data, ridge detection parameters including junction angles, backwards angle, start and stop thresholds were optimized with gradient descent or the simplex algorithm to achieve improved performance.

5.2.3 Feature Detection

Each point of membrane detected by the ridge detection method was then classified as synaptic profile, mitochondron, or normal membrane. Filter banks were created by rotating synapse and mitochondron weight matrices (Figure 5.2) to the same orientations to which the membrane Gabor-like patches were rotated. Note that twice the number of orientations were required because these weight matrices were not symmetrical.

At each point of detected membrane, synapse and mitochondria filter responses were calculated by element-wise multiplication and summation. Only two filter responses were calculated for each feature; one at the same orientation of rotation as the detected membrane, and one at the same orientation plus 180 degrees. Filter response thresholds were chosen to achieve acceptable error rates for synapse or mitochondria classification.

5.2.4 Edge Closure

Lines of membrane detected in this way usually failed to produce a fully segmented profile because line fragments near obliquely sectioned membrane remained unclosed. We were able to complete closure by identifying end points and joining them to neighbouring lines based on the shortest path through the energy function of the receptive field responses. Dijkstra's shortest path algorithm was used to calculate the shortest path over a 4-connected image graph based on the distance function shown in Equation 5.1, where $R_{x,\theta}$ represents the filter response at angle θ, centered at pixel x.

$$Dist_{xy} = max_{\theta}\left[R_{x,\theta}\right] * max_{\phi}\left[R_{y,\phi}\right] \tag{5.1}$$

For correctly detected edges, reasonable closure was achieved with this method, but incorrectly detected edges introduced additional edge errors when closing lines were added.

5.2.5 Alignment Improvement

Sequential images can be manually aligned by selecting several pairs of control points corresponding to the same x, y location for consecutive images z1 and z2. This approach is adopted by widely-used software, Reconstruct [28]. Automatic selection of control points is also possible by searching for unique image features in both images,

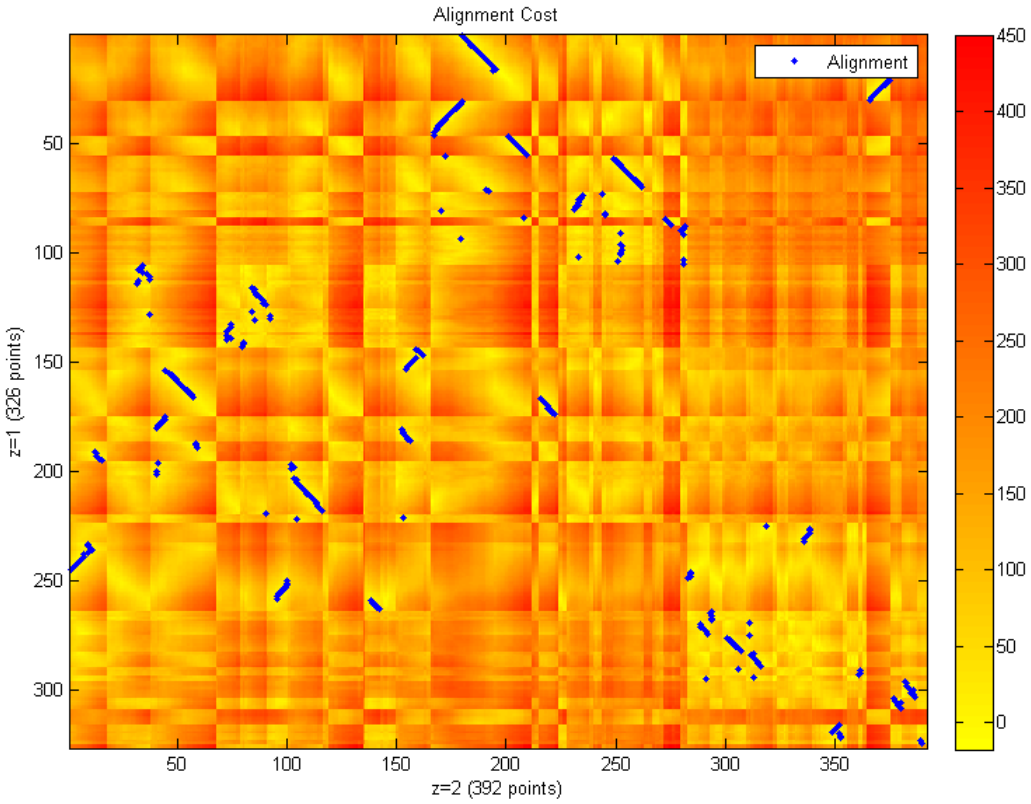


Figure 5.6: Alignment cost matrix obtained by calculating distance cost in Equation 5.6 ($a = 20$) for each combination of points between z1 and z2. Points of membrane from consecutive images z1 and z2 were ordered into two sequences preserving individual line segments, as described in Section 5.2.6. Each row corresponds to a point in z1 and each column corresponds to a point in z2. Cost matrix entries (orange to yellow) give the cost of matching points from z1 and z2. The best matching line segments appear as low cost (yellow) diagonal lines in this matrix. Blue dots highlight the resulting alignment after calculating similarity matrix H (not shown) and performing the traceback procedure. This alignment corresponds to the alignment shown in Figure 5.7 (middle).

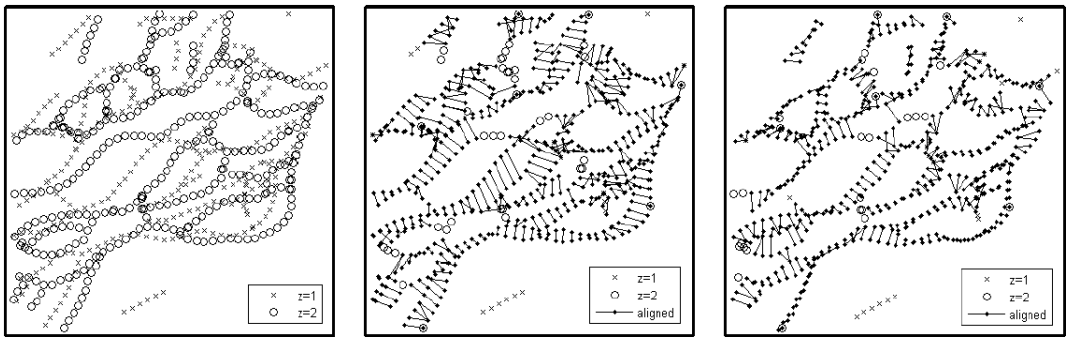


Figure 5.7: Alignment is improved by minimizing the average distance between all pairs of aligned points. Left: Start state. Points from z1(×) and z2(○) are shown before alignment. Middle: Alignment results based on Figure 5.6. Each pair of aligned points is shown as connected dots. Unmatched points remain as × or ○. A transformation was applied to z1 to minimize the distance between pairs of aligned points. Right: End state after repeated iteration of the alignment algorithm. Average directional offset between all aligned points is less than 1 pixel.

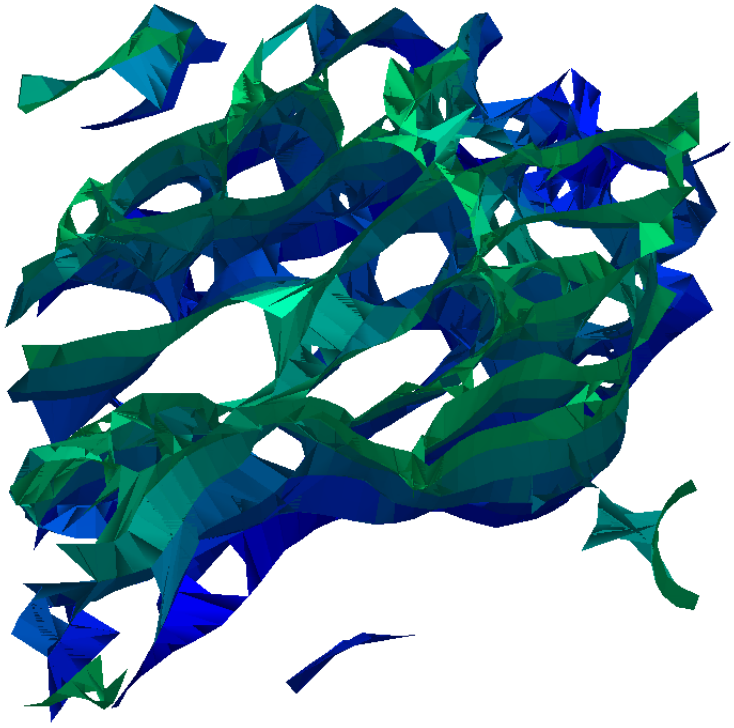


Figure 5.8: 3D rendering of membrane based on images from Figures 5.3 and 5.7. Segments of membrane aligned by the algorithm are represented as surfaces in this image. Six consecutive image patches were used. Surface colour (green to blue) represents z depth. Volume reconstructed is 1.4x1.4x0.3µm.

as demonstrated by software TrakEM2 [15, 115]. Automatic methods are usually effective at performing a global alignment, but significant local errors can be introduced when too few control points are detected, or when image features are inconsistent between images, as when sections have been locally distorted during microtomy or imaging, resulting in the need for manual correction [15].

Lines of detected membrane are a useful means to improve an existing image alignment. Because membrane is abundant in EM images many potential control points can be found by aligning ridge detection results. Drawing on experience with the linear alignment of other biological structures for this purpose, we therefore developed a dynamic programming algorithm, similar to the Needleman-Wunsch [89] and Smith-Waterman [122] DNA sequence alignment procedures, with a cost metric based on the euclidean distance and angular subtense.

The algorithm also has similarities with sequence matching algorithms implemented in 2D for curve morphing [52] and in 3D for neuron shape recognition [14]. We introduced a different cost metric, and a modified 3-pass alignment procedure that could perform many-to-many alignments and allows branching to occur within sequences. Combinations of this new alignment procedure with the existing application areas allowed morphing of multiple curves at once, and recognition of branching neuron shapes or even networks of branching neurons.

Two consecutive images (z1, z2) from the image stack were aligned by matching sections of detected membrane in z1 with sections in z2, so that the distance and angular subtense between all matched points were minimized. This problem was similar to a many-to-many ends-free DNA sequence alignment with the cost metric shown in Equation 5.2, where $d(p_1, p_2)$ is the euclidean distance between points p_1 and p_2, and a is an arbitrary angle constant ($a = 20$ for our implementation). In principle, alignment can occur in either the forwards or backwards direction, so that low-cost diagonal lines in the cost matrix indicate the best alignment. Similarity matrix H was calculated and the traceback procedure [89] used to find the best alignment of points in z1 to points in z2. An example of a cost matrix and the corresponding alignment points is shown in Figure 5.6. Further details, including the calculation of the similarity matrix, a described in Section 5.2.6.

$$Cost\,(p_1, p_2) = d\,(p_1, p_2) + a|p_{1\theta} - p_{2\theta}| \tag{5.2}$$

Once alignment was complete, the average offset between aligned points was calculated and used to improve alignment between z1 and z2. This process was repeated

until the average directional offset was less than 1 pixel. The result from a single alignment is shown in Figure 5.7.

This alignment method assumes that the direction of membrane movement between consecutive images, when averaged over a sufficiently large area, is close to zero. For example, in a given alignment the amount of membrane moving to the left is assumed to be approximately equal to the amount of membrane moving to the right. Depending on the angle of ultrathin sectioning and the particular area of neural tissue being imaged, it is possible that there will be a bias in the direction of overall membrane movement. This bias may exist for the entire image, or for small sections of it, especially where there are large bundles of neurites all running in the same direction.

Image alignment in this case does not differentiate between distortions due to preparation or imaging artifacts and areas of bias resulting from membrane movement. Ideally we would like to correct for distortions while preserving any movement bias. However, based on 2D control points alone this problem is ambiguous without further information.

One solution to this ambiguity is to perform membrane alignment only on the highest-scoring sections of membrane; these sections of membrane have clearer, thinner profiles in the image and are expected to be perpendicular to the cutting plane. In this way we can assume that any alignment errors are more likely to be from distortions rather than membrane movement, and can use a deformation transform with greater confidence. However, this method can still introduce small errors that accumulate, resulting in large errors over many sections.

A second solution is to simply use linear translation and rotation for the alignment to ensure that no unwanted distortions are introduced. This method may result in poor alignment where there are large areas of imaging artifacts.

Alignment results are considered as the surfaces of cell membranes or organelles in 3D. By combining multiple alignment results it was possible to generate partial 3D reconstructions as shown in Figure 5.8.

5.2.6 Alignment Algorithm

The following dynamic programming algorithm was developed to improve alignment between membrane detected in slice z1 to membrane detected in consecutive slice z2. The procedure is similar to the Needleman-Wunsch [89] and Smith-Waterman [122] algorithms for DNA sequence alignment. This algorithm can be used to align

any two segmentations or edge detection results provided that edges are broken into a finite number of points with approximately equal spacing and each point can be associated with an angle of orientation. Approximate alignment between images z1 and z2 should be performed first by some other control point detection method. To reduce computational complexity, large images should be broken up into smaller image patches and alignment repeated separately for each pair image patches (512x512 pixels was a reasonable patch size for our images, as shown in Figure 5.7).

Let points of membrane detected in image z1 be represented as by sequence A, consisting of a set of line segments, $S_1 \ldots S_n$, each of which is an ordered list of connected points, $S_a = \{p_{a,1} \ldots p_{a,m}\}$. Each point is located in a 2d euclidean plane with coordinates x, y, along with an associated angle of maximum filter response θ. For convenience, A is also represented by the ordered list of points $p_{1,1} \ldots p_{n,k}$ directly:

$$\begin{aligned} A &= \{S_1, \ldots, S_n\} = \{p_{1,1}, \ldots, p_{1,m}, p_{2,1}, \ldots p_{n,k}\} \\ S_a &= \{p_{a,1}, \ldots, p_{a,m}\} \\ p_{a,b} &= \{x, y, \theta\} \end{aligned} \tag{5.3}$$

Membrane is split into line segments $S_1 \ldots S_n$ so that all start points $p_{a,1}$ and end points $p_{a,m}$ are junction points or end points. Additionally no junction points exist between $p_{a,1}$ and $p_{a,m}$. Junction points are added multiple times, once for each line segment that connects to it. Thus each line segment S_a is a one dimensional sequence of connected points. The order of line segments $S_1..S_n$ is not important, but points $p_{a,1}..p_{a,m}$ are ordered according to the line segment connectivity. For convenience, sets $A_{start\_points} = \{p_{a,1}; a = 1..n\}$ and $A_{end\_points} = \{max_k(p_{a,k}); a = 1..n\}$ are defined to contain all start and end points of sequences $S_1 \ldots S_n$ respectively. Sequence B is similarly constructed from membrane detected in consecutive image z2.

To align points in sequence A with points in B we construct the similarity matrix H, with rows of H (indexed by i below) corresponding to ordered points from A and columns of H (indexed by j below) corresponding to ordered points from B. Note that this alignment is asymmetric. All points in A are aligned with the best candidate points in B, or remain unmatched with a gap penalty, so that the alignment score is maximized. Some points in B may remain unmatched without affecting the score.

Columns of H are filled in over three passes. In the first pass rows are visited in ascending order of j considering $match diag down$ (:mdd), $match across$ (:ma), $gap across$ (:ga), and $gap down$ (:gd) moves. When i represents the start point of a line segment in A all scores in the column are set to 0. When j represents a point at the start of a line in B

only $gapacross$ (:ga) moves are considered in the first pass. (5.4)

Then in the second pass, rows are visited in descending order of j, considering the best move from the first pass (:nc), $gapup$ (:gu) and $matchdiag\_up$ (:mdu) moves. When j represents a point at the end of a line in B, no moves are considered in the second pass. (5.5)

The third pass identifies possible $split$ (:sp) moves, for a line in A that matches more than one line segment in B, possibly corresponding to a junction point or a broken line in B. Any j that represents a start or end of a line segment in B ($B_{start\_points} \cup B_{end\_points}$) is a candidate split point. The highest scoring candidate split point in column i after pass 2 is identified and the score of this entry is recorded as $split\_score$, with the location recorded as $split\_j$. Any other candidate split point in column i with a score of less than $split\_score - split\_cost$ is updated to have this score. (5.6)

For each move, the match boolean and previous point location are recorded as summarized in (5.7) so that the traceback procedure can be used and matching points identified. Note that the previous step from a split move updated in phase 3 is $(i, split\_j)$.

Pass 1:

$$H(i,j) = \ 0 \quad \text{:s} \ \forall \left\{ i \in A_{start\_points} \right\}$$

$$H(i,j) = \ H(i-1,j) - gap\_cost \quad \text{:ga} \ \forall \left\{ \begin{array}{c} i \notin A_{start\_points}, \\ j \in B_{start\_points} \end{array} \right\}$$

$$H(i,j) = max \left\{ \begin{array}{cc} H(i-1,j-1) + match - Cost(i,j) & \text{:mdd} \\ H(i-1,j) + match - Cost(i,j) & \text{:ma} \\ H(i-1,j) - gap\_cost & \text{:ga} \\ H(i,j-1) - gap\_cost & \text{:gd} \end{array} \right\} \forall \left\{ \begin{array}{c} i \notin A_{start\_points}, \\ j \notin B_{start\_points} \end{array} \right\} \tag{5.4}$$

Pass 2:

$$H(i,j) = max \left\{ \begin{array}{cc} H(i,j) & \text{:nc} \\ H(i-1,j+1) + match - Cost(i,j) & \text{:mdu} \\ H(i,j+1) - gap\_cost & \text{:gu} \end{array} \right\} \forall \left\{ \begin{array}{c} i \notin A_{start\_points}, \\ j \notin B_{end\_points} \end{array} \right\} \tag{5.5}$$

Pass 3:

$$H(i,j) = max\left\{\begin{array}{cc} H(i,j) & :\text{nc} \\ split\_score - split\_cost & :\text{sp} \end{array}\right\} \forall \left\{\begin{array}{c} i \notin A_{start\_points}, \\ j \in \left(\begin{array}{c} B_{start\_points} \\ \cup B_{end\_points} \end{array}\right) \end{array}\right\} \tag{5.6}$$

Move Summary

Move	Pass	Direction	Match	Previous
:s = *start*	1		0	$\emptyset$
:mdd = *matchdiagdown*	1	$\searrow$	1	$i-1, j-1$
:ma = *matchacross*	1	$\rightarrow$	1	$i-1, j$
:mdu = *matchdiagup*	2	$\nearrow$	1	$i-1, j+1$
:ga = *gapacross*	1	$\rightarrow$	0	$i-1, j$
:gd = *gapdown*	1	$\downarrow$	0	$i, j-1$
:gu = *gapup*	2	$\uparrow$	0	$i, j+1$
:sp = *split*	3	$\Updownarrow$	0	$i, split\_j$
:nc = *nochange*	2,3			

(5.7)

The cost between two matching points p_1 and p_2 was chosen to minimize the overall euclidean distance $d(p_1, p_2)$ and angular subtense between the two points (5.2). Angle constant a can be chosen to give more or less importance to the angular subtense.

In our implementation the cost constants shown in (5.8) were chosen to reward alignments with fewer gap or split moves. These constants can be modified to achieve different alignment properties.

$$\begin{aligned} match &= 20 \\ gap\_cost &= 20 \\ split\_cost &= 30 \\ a &= 20 \end{aligned} \tag{5.8}$$

Matched points are identified by finding the highest scoring point for each column i in A_{end_points} and using the traceback procedure to trace the alignment through the matrix until the corresponding start point is reached. Any *matchdiag_down*, *matchacross*, or *matchdiag_up* moves indicate the matching points between points A_i and B_j.

Once all pairs of matching points have been identified (for all image patches) they are treated as control point pairs. A 2D transformation is constructed to minimize the

distance between all control point pairs and this transformation is applied to image z1 and the corresponding points of membrane in A. The alignment algorithm is then repeated until the average distance between control point pairs is acceptably small. In principle any deformation transformation could be used to minimize the distance between control points, however we chose to use a rotation and translation transform so that local shapes and sizes were preserved. Deformation transformations would be more appropriate where there is evidence of local distortion in the original image.

5.2.7 Results

Manual reconstructions of EM data are difficult to compare directly with segmentations derived from algorithms. Reconstructions are usually performed with the intention of tracing the neuron correctly over many sections rather than identifying the exact location of the cell membrane in every image. With these goals in mind, manual reconstructions generate the general shape of the neuron and overall neuron morphology along with the contacts made with other neurons correctly, but with the exact location of membrane not necessarily accurate in all places. In areas where a membrane runs obliquely in the section and appears blurred in its corresponding projection image, or where a large presynaptic density is present (Figure 5.1b, 5.1c), membrane signal can occupy a width of 20 pixels or more at a resolution of 3.7nm per pixel. For the same data, tracing variation between experts can be up to 20 pixels, or 74nm, depending on both the acceptable level of accuracy of tracing with a manually controlled mouse, and true uncertainty in the location of oblique membranes (data not shown). This looseness in manual tracing makes direct comparison between manual and automatic tracing methods difficult to achieve. Choosing a performance metric that recognizes topological correctness rather than small differences in boundary locations [48] and using high quality image data against which to assess automatic tracing are both important considerations.

To overcome the problem of the disparity between manual and automatic tracing methods, an interactive web interface was developed to view and correct membrane automatically traced from EM images that had previously been annotated manually. Further details of the annotation system are included in Section 5.3. Errors made by the algorithm were classified as either false positives (locations where membrane was detected by the algorithm but was not actually present) or false negatives (where membrane was present but not detected). Using the web interface we identified false posi-

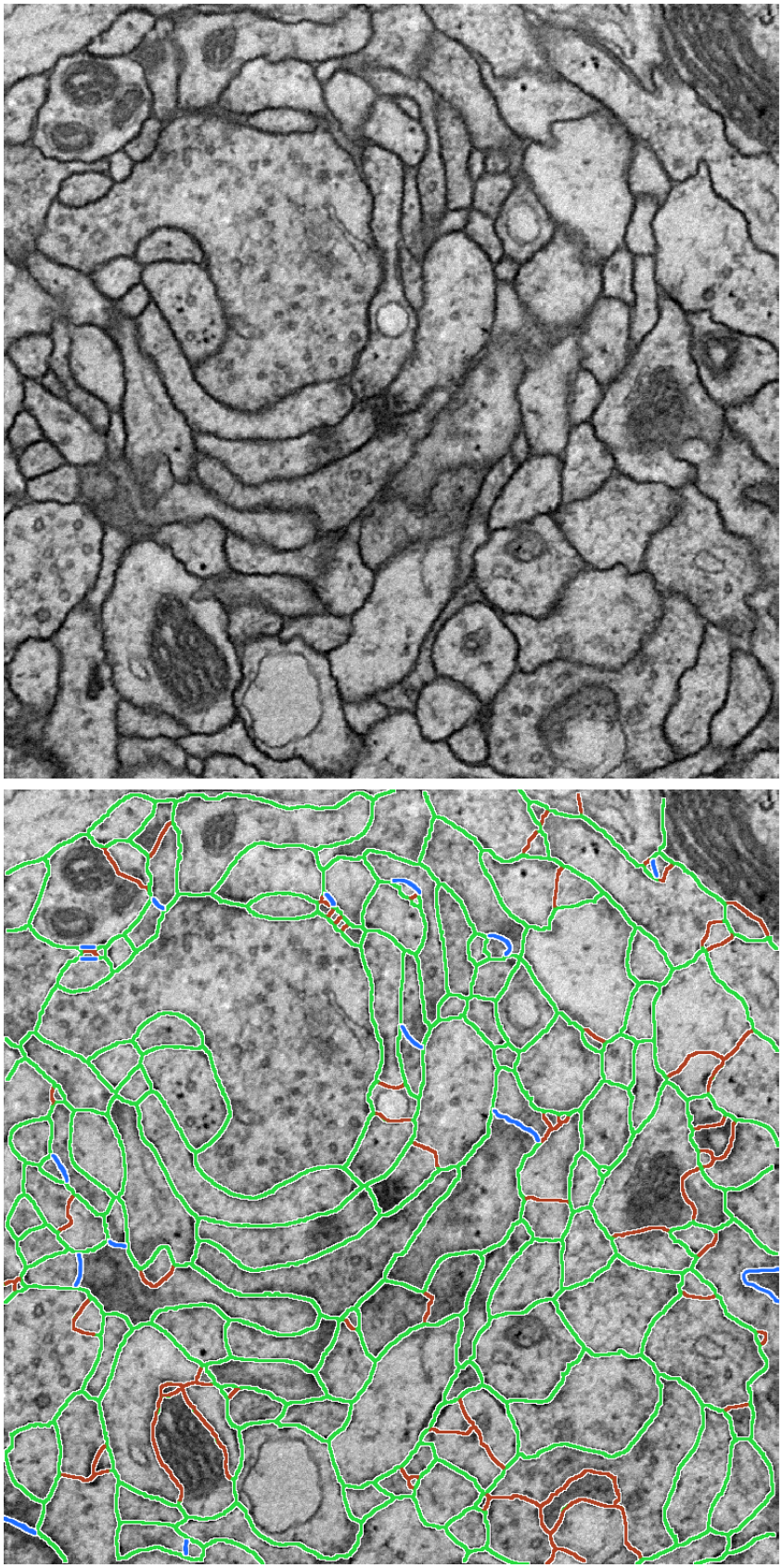


Figure 5.9: Segmentation evaluation. Top: Test EM image. Bottom: After ridge detection, edge closure and manual correction (as described in Section 5.3). True positives (green, 97.4%) false positives (red, 14.9%), false negatives (blue 2.6%). Image size is 2.8x2.8µm.

tive lines by clicking on them, and drawing in manually the missing, or false negative, lines. Using this method all errors were identified and a fully traced membrane data set was constructed within a small volume.

Selected trace results and alignment improvements were imported into manual reconstruction software, Reconstruct [28], for direct visual comparison with manual tracing. 3D renderings of results are shown in Figure 5.10. Alignment improvement and semi-automatic tracing produced a more accurate representation of the reconstructed bouton of a projection neuron, the main input neuron to the mushroom body calyx. The semi-automatic annotation is smoother and small misalignments in the z direction are corrected. The correlation coefficient between pairs of consecutive images was also calculated for the volume shown in Figure 5.10. The average correlation coefficient was 0.29 after manual alignment, and 0.32 after alignment improvement using a linear transformation. Note that this level of accuracy can also be achieved by careful manual annotation but would take much longer to complete. Exact membrane accuracy is usually traded for faster, less accurate tracing that preserves topological correctness.

The fully traced membrane data was used to optimize and test algorithm performance. Convolutions necessary for the line detection algorithm were implemented on a graphics processing unit (GPU) to improve algorithm speed. Edge detection parameters were first estimated empirically, and then optimized by simplex or gradient descent optimization to maximize metric scores. The Rand index, a commonly used measure of segmentation performance [106, 133, 130], was used to assess performance, as shown in Table 5.1. Performance was also measured by the number of separating pixels between segments that were correct or incorrect as a proportion of total true positive separating pixels, as shown in Table 5.1 and displayed in a receiver operator characteristic (ROC) in Figure 5.11. Membrane detected within 10 pixels of manually annotated membrane was considered correct, because for the ssTEM images used here, this width would correspond to a flat section of membrane at an oblique angle of 36 degrees.

Performance was benchmarked against GSH [88, 134] and a freely available random forest classifier, ilastik [20], manually trained on a range of generic features to identify cell membrane. Scores from both these benchmarks were segmented by the watershed algorithm. Parameters were optimized to maximize metric scores in both cases.

Responses of the Gabor-like receptive fields were robust to several types of noise sometimes encountered in ssTEM images such as low contrast images, blurred or out-of-focus areas, and sudden or gradual changes in brightness. After optimization the

	Rand Index	Tp	Fp	Fn
Open Edges	0.791	0.976	0.254	0.024
Closed Edges	0.762	0.978	0.282	0.022
Ilastik / Watershed	0.711	0.973	0.283	0.027
GSH / Watershed	0.718	0.948	0.521	0.052

Table 5.1: Membrane detection performance: Rand index is expressed as a measure of similarity, with 1 being identical to the manually corrected segmentation. Separating pixel true positive (Tp) false positive (Fp) and false negative (Fn) rates are shown as a proportion of the total true positive separating pixels. Algorithm parameters were optimized by simplex or gradient descent to find approximately 10 times more false positives than false negatives or to maximize the Rand index score. 5-fold cross validation was used to validate Rand index scores.

ridge detection and edge closure methods were able to join gaps where noise such as oblique membrane or stitching artifacts obscured the receptive field responses.

Line segments identified by edge detection and edge closure operations were further classified as enclosing the profile of either a synapse or a mitochondrion by the receptive fields shown in Figure 5.2a. Feature detection performance is shown in Figure 5.12. False positive rates were higher than those for membrane detection, however many false positives were identified in regions near an actual synapse or mitochondria. This level of performance could be useful for narrowing down search areas for manual classification of such biologically significant features.

We also trained the ilastik classifier using the membrane receptive field responses as input features. When trained on receptive field responses alone, results were slightly better than those when trained on generic features (0.72 Rand index). Results improved further when trained on both receptive field responses and generic features (0.74 Rand index).

Results are still far from perfect, and errors in each 2D image propagate in 3D to completely obfuscate the reconstructed connectivity network. However, when combined with a usable computer interface, the imperfect automatically generated annotations can be used as a starting point to reduce manual annotation time required to make a complete annotation. In the next section I describe a basic web-based annotation system developed to generate the gold standard annotations used here as training as test data.

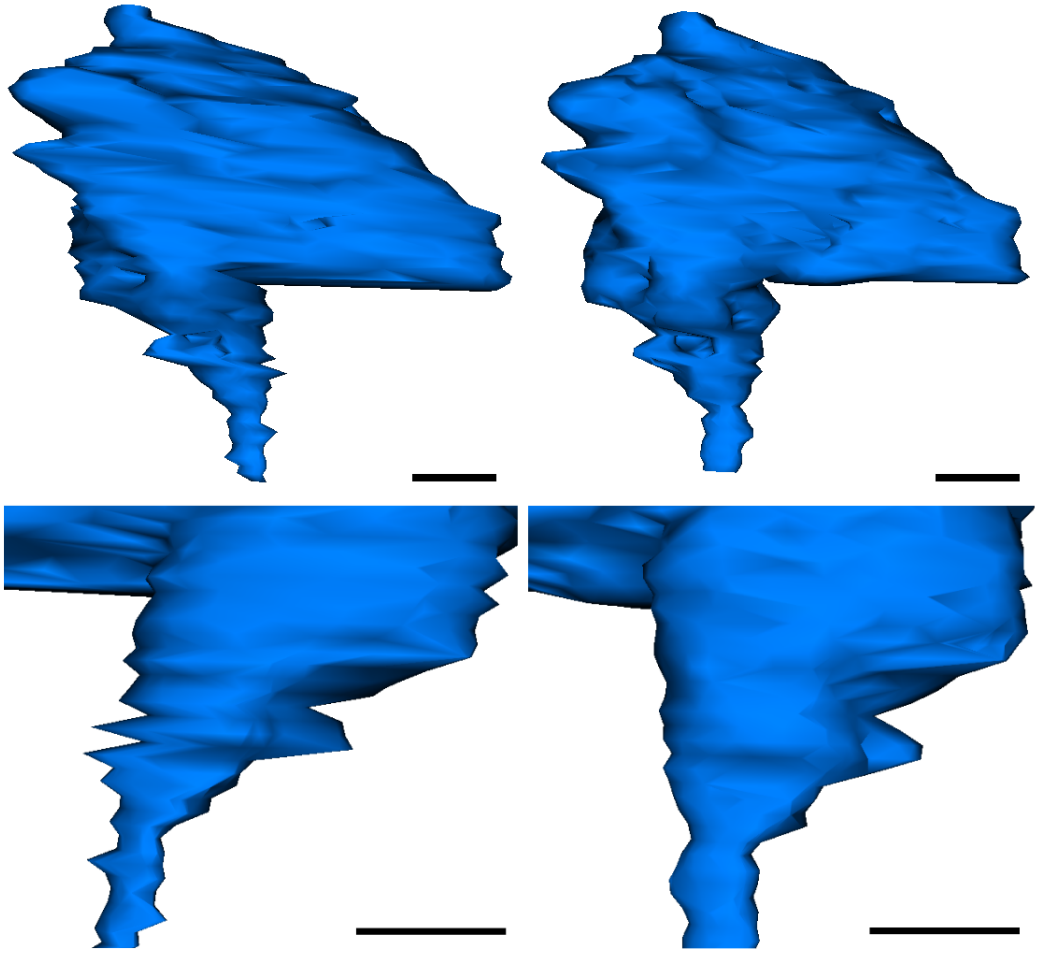


Figure 5.10: 3D renderings of a reconstructed bouton and axon of a mushroom body calycal projection neuron. Left: Manual alignment and tracing. Right: The same volume after alignment improvement and semi-automatic tracing. Small alignment errors are improved, resulting in a less jagged surface profile. Scale bars: 0.5µm.

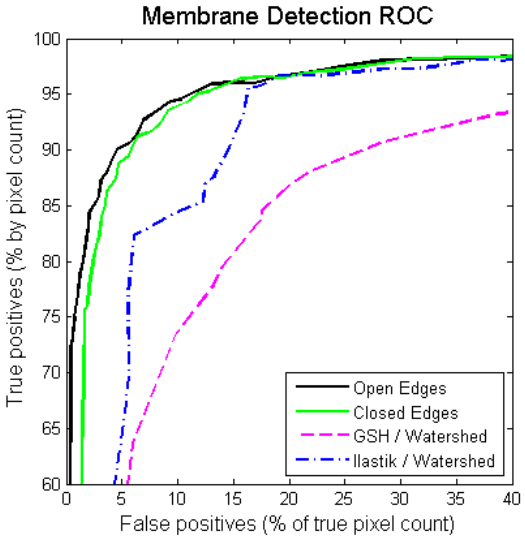


Figure 5.11: Receiver operator characteristic (ROC) curve for membrane detection performance after optimization. Results were assessed before and after edge closure (Open Edges and Closed Edges respectively). Performance is compared against the watershed algorithm applied to an optimized Gaussian smoothed Hessian (GSH), and to a manually trained random forest classifier (ilastik). False positives are expressed as a percentage of true separating pixels, as determined by manual annotation.

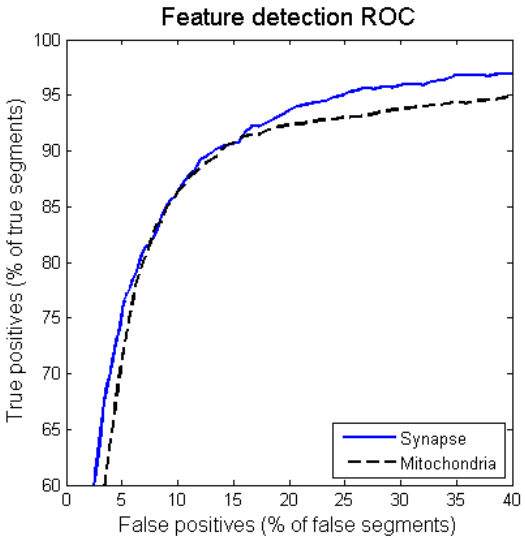


Figure 5.12: ROC curve for feature detection performance of synapse and mitochondria profiles. Line segments found by edge detection and edge closure were classified as either synapses or mitochondria, and results were manually corrected. False positive rates are expressed as a percentage of line segments not in the target class (line segments that do not form part of any synapse or mitochondria respectively).

5.3 Density Website

Because no algorithms yet exist for completely automated reconstruction of ssTEM images, it is necessary for images to be manually annotated. Fortunately, using computer-based segmentation as a starting point it is much faster for an annotator to correct the errors of an automatic segmentation than it would be to annotate the volume from scratch. To generate training and test data used in [62], I developed a simple online annotation system, named Density, for correcting automatic ridge detection results (Figure 5.13). A demonstration system is available at http://fruitfly.inf.ed.ac.uk/density/.

Controls at the top of the screen allow the user to toggle the detected membrane overlay, zoom in an out, and move up and down in the image stack (Figure 5.14).

To annotate images the user first chooses an initial confidence level to select the approximate amount of membrane enabled (Figure 5.15 A). Disabled membrane appears faded and enabled membrane appears green according to algorithm confidence level. After choosing an initial confidence level the user can click on detected membrane to change it from enabled to disabled and back again (Figure 5.15 B). Missing sections of membrane, not detected by the algorithm, can be drawn in manually with a series of clicks on the image (Figure 5.15 C). By repeating this procedure throughout the image it is possible to correct the automatically detected membrane (Figure 5.15 D).

The Density annotation system can also be used to annotate synapses or mitochondria, and was used to provide a training and test data sets for assessment of feature detection performance.

5.4 GPU Implementation

Recent advances in graphics processor technology, driven by the gaming industry, has driven development towards many cores on a single graphics processing unit (GPU). We can use these cores to perform computations and increase processing speed by up to 100 times [96, 91]. However, limitations on memory transfer bandwidth and algorithm implementation details mean that speedup is not always 100 times. I found that 35 times speed up was possible for convolution operations used for neuron tracing in section 5.16. This kind of speedup was not possible previously without cluster or super computer implementations, so GPU technology offers a very cost effective way to perform computationally expensive operations.

3D volumetric data is difficult to analyse computationally because processing time

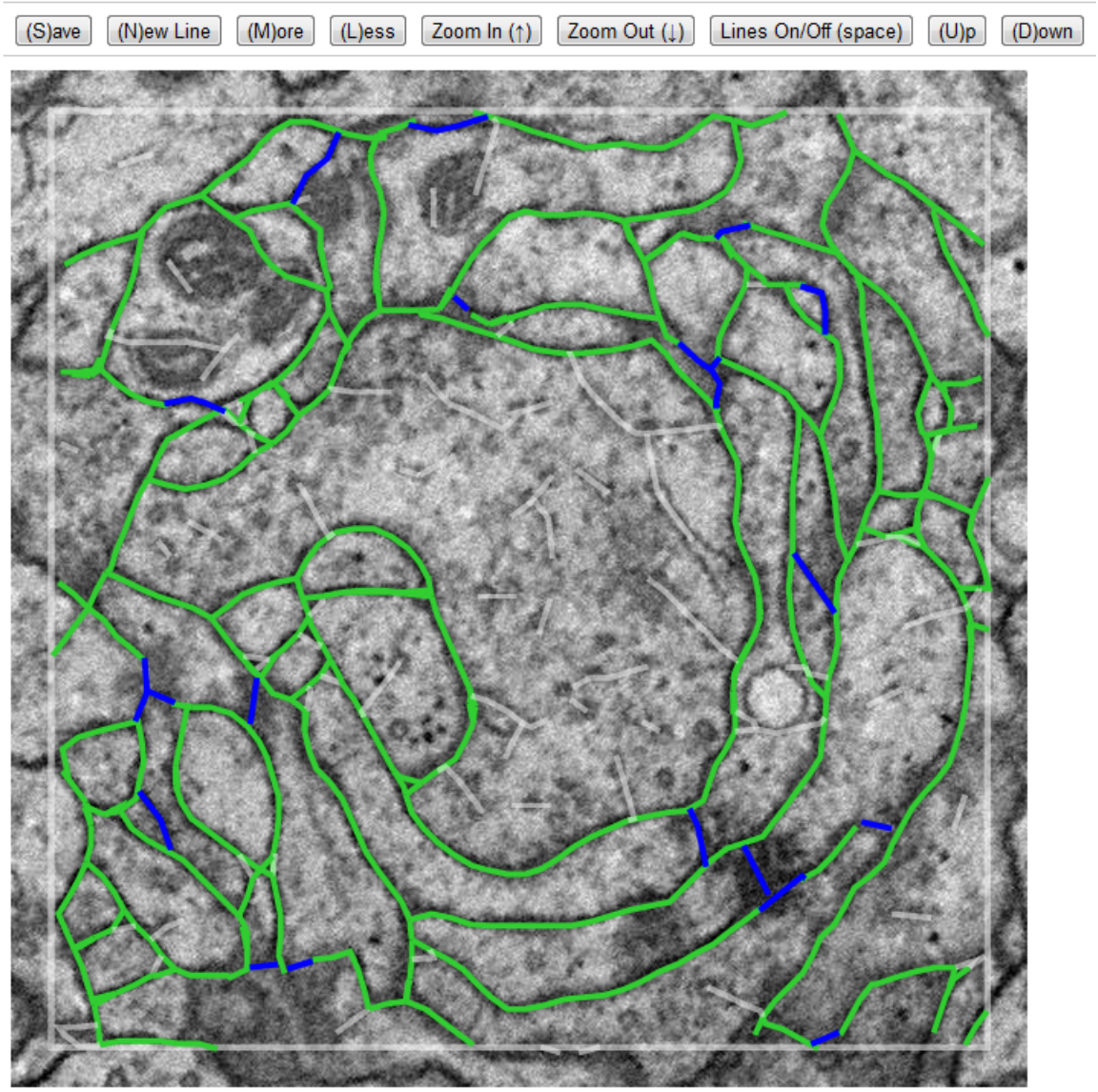


Figure 5.13: Density website interface. In this view detected membrane edges are overlaid on top of the EM image. This image has been corrected manually, green lines indicate true positive detections, gray lines false positives, that were manually deselected, and blue lines false negatives, that were manually added.

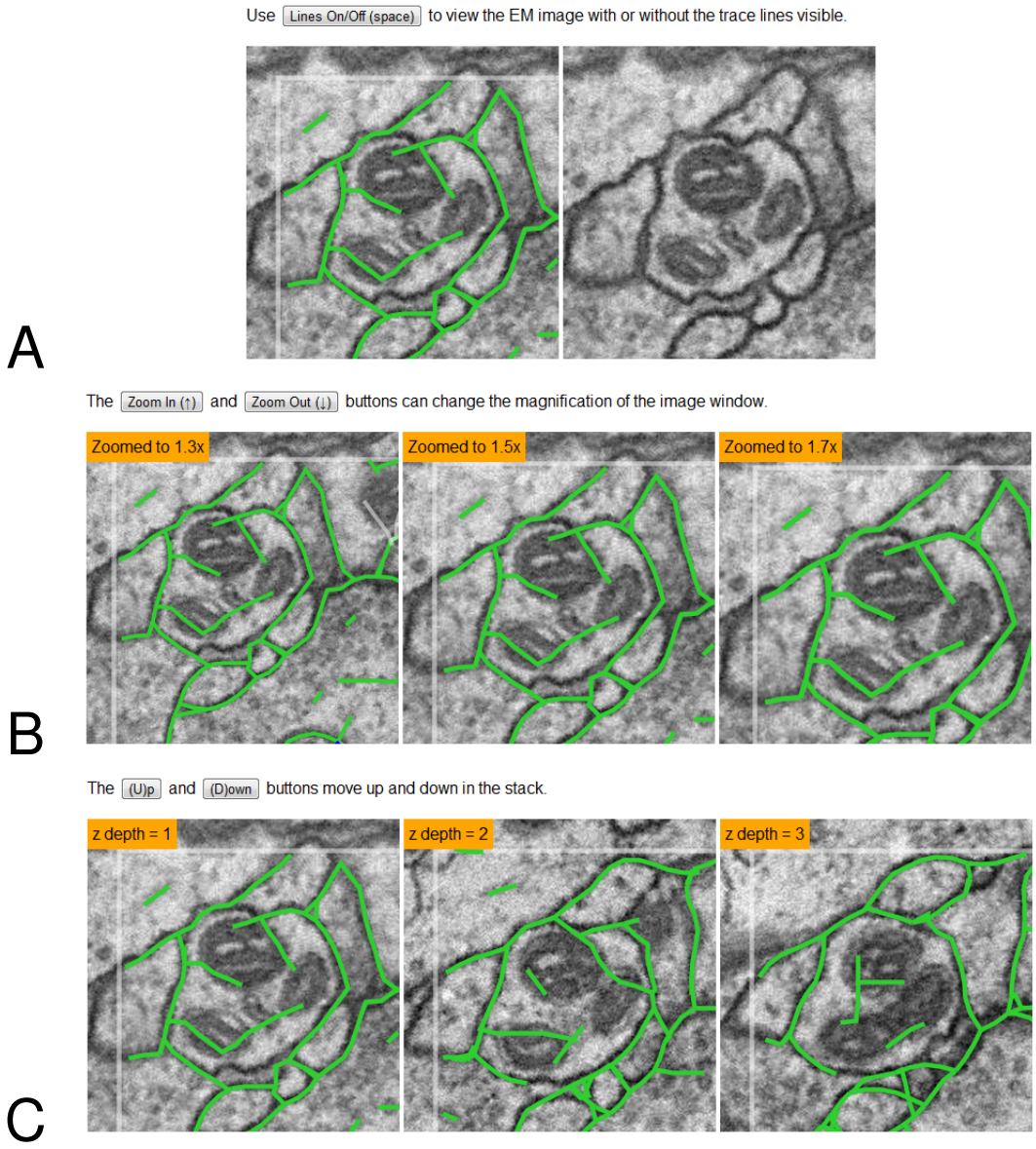


Figure 5.14: Density view controls. The user can toggle overlay lines on and off (A), to view the EM image directly, zoom in and out (B), and navigate up and down in the image stack (C).

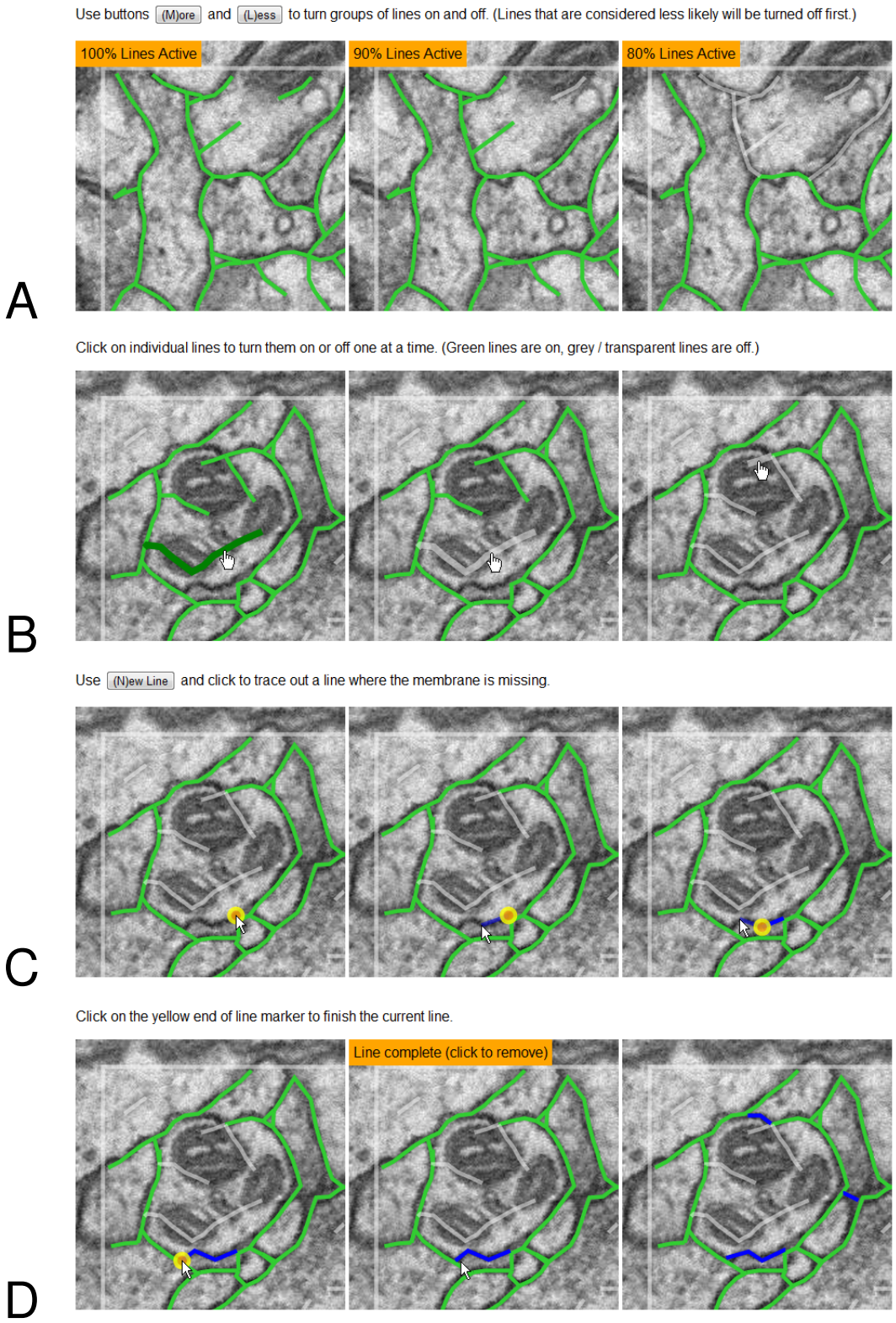


Figure 5.15: Density trace correction. Using the web interface the user can automatically enable or disable lines based on the algorithm confidence (A), enable and disable individual lines by clicking on them (B), and draw in additional lines of membrane not detected by the algorithm (C,D). Manual correction is complete once all lines have been toggled to the correct state, and extra lines added (D, right).

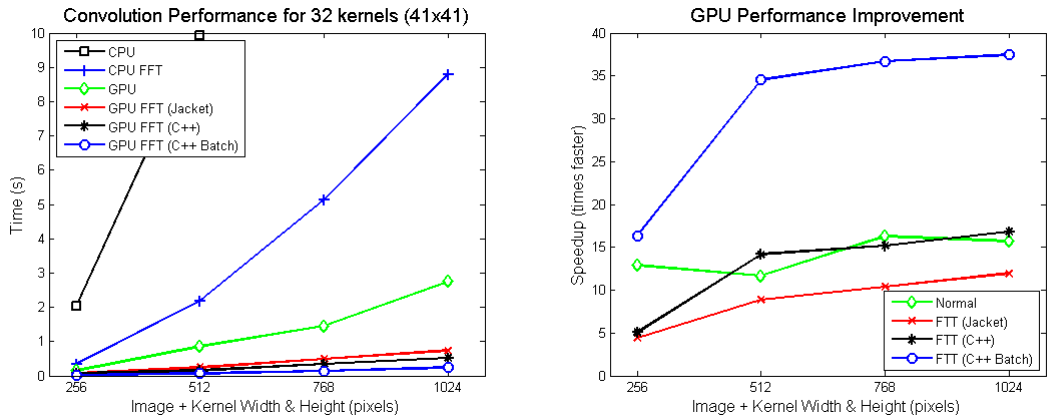


Figure 5.16: Comparison of 2D convolution performance between a 4-core desktop CPU (Intel core i5 750) and a 240 core GPU (nVidia GTX 285). Conventional and Fast Fourier Transform (FFT) convolution methods are compared.

required for each operation quickly becomes prohibitive for large volumes. I have implemented convolution-based 2D tracing algorithm operations for EM images, as mentioned previously, and extended receptive fields to 3D convolution for near-isotropic EM images. Convolution operations, that form the basis of many algorithms, can be efficiently implemented on GPU systems. Figure 5.16 shows performance improvements attained so far for 2D convolution, which forms the basis of tracing algorithms developed for EM image analysis.

In the next section, I describe an extension of the 2D ridge detection system with 3D receptive fields. Applied to publicly available FIB/SEM data from Graham Knott, this system can achieve 88.84% segmentation accuracy (Rand index).

5.4.1 3D Receptive Fields

One way to improve the performance of the tracing algorithm is to improve the filter accuracy by creating 3D filters. This presents two problems for us to solve. Firstly, the amount computation required for convolution in 3D is an order of magnitude greater than for 2D, making many experiments very time consuming to perform on a normal computer CPU. This problem is solved by using GPU computing, which allows speedup of around 35 times in this case. The second problem is that in 3D there are many more membrane surface shapes possible, and the type of filters used become even more important. I address this problem by adapting 2D filters to the 3D environment, and curving them to match several possible shapes of small patches of membrane in 3D.

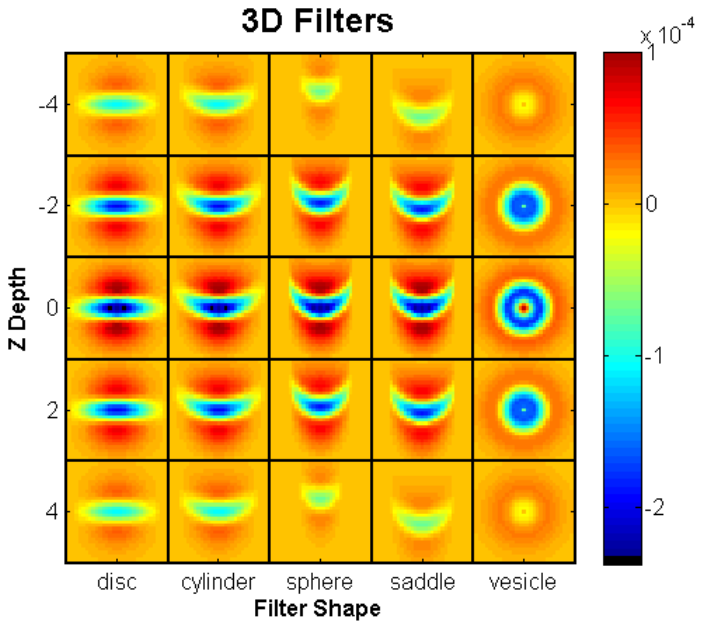


Figure 5.17: 3D receptive fields based on 2D receptive fields shown in Figure 5.2. Receptive fields were rotated to form a disc shape in 3D, and curved to approximate possible membrane and vesicle shapes observed in FIB/SEM data. Full receptive fields consist of 15 z slices; 5 slices are shown. Pixel colours represent weights from input pixels, as shown in the right hand key.

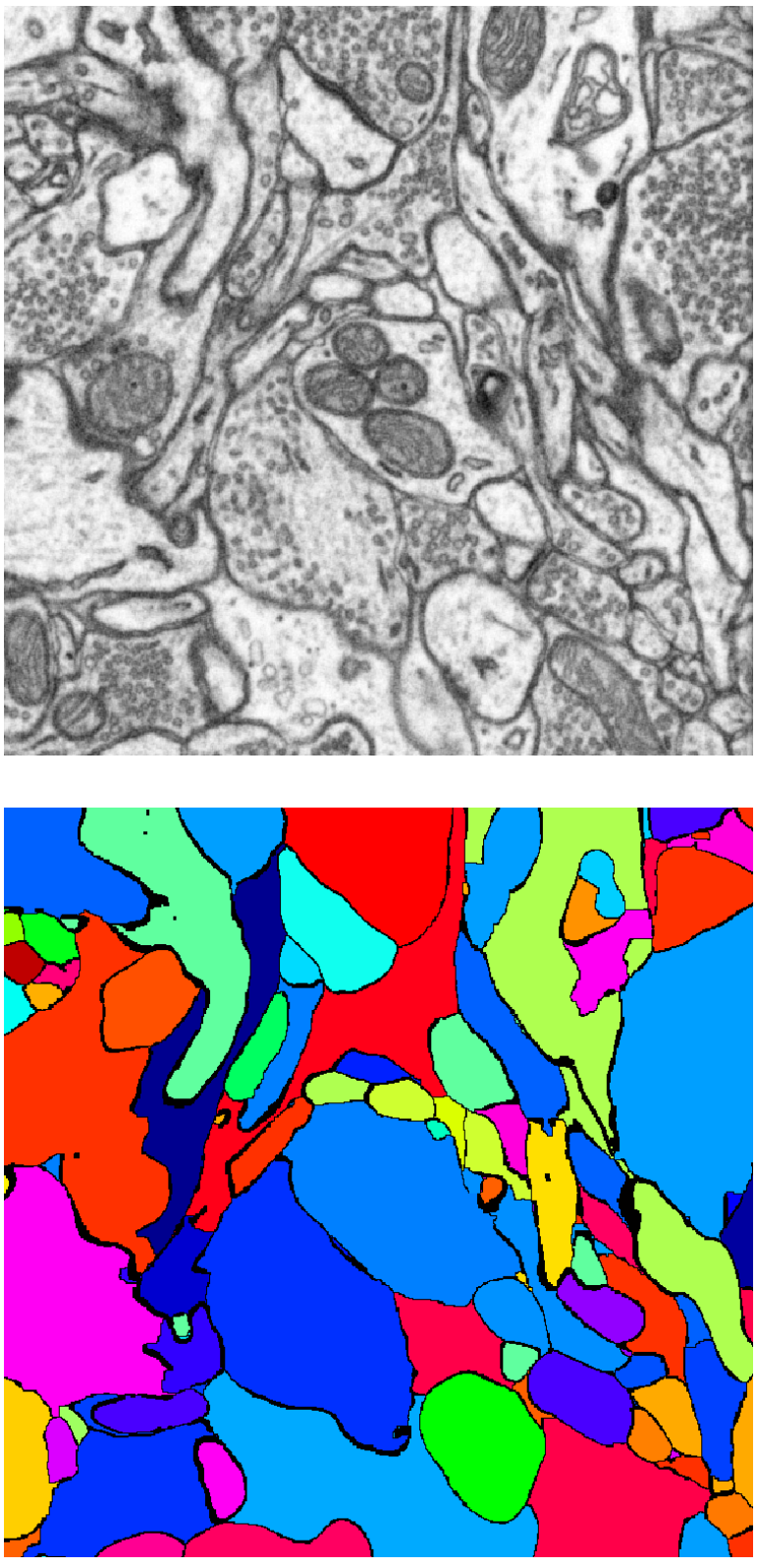


Figure 5.18: Segmentation results from 3D receptive fields based on publicly available FIB/SEM data provided by Graham Knott, EPFL. Top: Original image. Bottom: Segmented image, false coloured to show different neurons. The segmentation shown achieves a Rand index (2D) of 82.79%. Each image is approximately 3.2x3.2µm

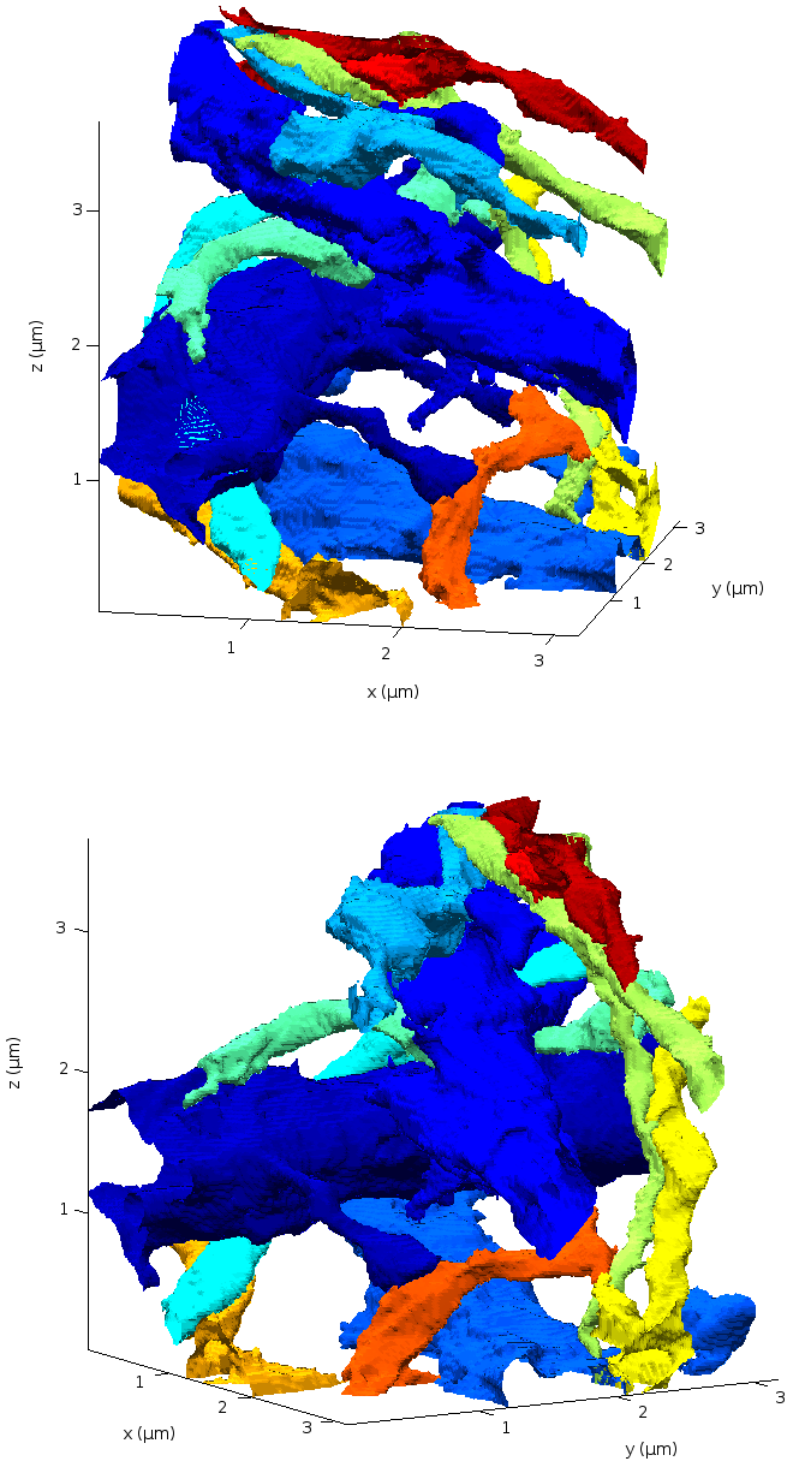


Figure 5.19: Automatic segmentation results. 12 segments are shown from a small volume. The same 3D scene is shown in both images, slightly rotated. Neurites oblique to the imaging plane are segmented.

2D receptive fields generated in Section 5.2.1 were rotated to form disc shaped receptive fields, and warped to form cylinder, ellipsoid, and saddle shaped receptive fields (Figure 5.17). Warping receptive fields further to form small spherical shapes generated receptive fields were used to detect vesicles. A selection of disc, cylinder, ellipsoid and saddle receptive fields were rotated to 300 different angles (evenly distributed around a sphere) and convolved over an input 3D image.

The anisotropic nature of ssTEM images made them difficult to analyze with 3D receptive fields. For further investigation into this method I used publicly available data, provided by Graham Knott, EPFL, of neurons from rat striatum, imaged in a focused ion beam (FIB) / Scanning Electron Microscope (SEM). This method of EM imaging produces near-isotropic volumetric images; in this case a resolution of 5x5x9nm per voxel was achieved.

After convolution, a combination of ridge detection (in 2D), as described in section 5.2.2, and the 3D watershed algorithm was used to segment images. The initial segmentation was cleaned up by joining any small segments below a chosen size threshold to their neighbouring segments. Assessed on a small volume, a Rand index performance of 88.84% was possible with this system (Figures 5.18 and 5.19). This shows a significant improvement over the 2D ridge detection from Section 5.2.2, which achieved a Rand index score of 68.26% assessed on the same data, without any optimization.

It is not yet possible to apply this approach to all types of EM data, however it is promising that GPU technology and new imaging techniques can improve automatic EM segmentation performance significantly.

5.5 Summary

This chapter presented a set of computational methods for EM image alignment and reconstruction, based on a set of receptive fields learnt from EM image data. The identification of many control points for aligning consecutive images can improve upon manual alignment methods and is robust to many types of noise encountered in EM images. Closing edges based on a shortest path algorithm can also achieve a full segmentation of images, and additional receptive fields can be used to identify the profiles of synapses and other organelles present in ssTEM images.

The ridge detection approach is complementary to existing regional or watershed-based methods, and achieves similar or superior results. Aligning points of membrane

by the dynamic programming algorithm is also complementary to existing control-point based alignment methods and can improve upon these in some cases, especially at places where areas of noise or imaging artifacts affect control point properties.

The GPU implementation of the tracing system using 3D receptive fields showed promising performance improvements, and this area offers the potential for future research. Improving 3D receptive fields, and the combined ridge detection and watershed system is one area for further study, and further research into ssTEM image processing could also benefit from 3D receptive fields.

The manual alignment and segmentation of detailed ssTEM images is very time consuming, but information on synaptic connections obtained by these means is essential for research in systems neuroscience [10]. This reawakened need has recently received renewed recognition, identified in the recently designated field of connectomics [74]. Inspired by the example of tools used in biology for molecular alignment, the set of methods we report for improved alignment and detection of membrane is able to assist in the time-consuming process of manual annotation.

Further information about likely membrane locations is also available from consecutive images in the stack. Areas where membrane alignment is poor between two images in the z axis may indicate a false positive or false negative identification in either image. Utilization of this additional information and further improvements in both image processing techniques and image quality will help lead to the complete automation of neuronal reconstruction in 3D, and the complete identification and definition of circuits constituted by such reconstructed neurons.

Currently all tracing algorithms described here are implemented in Matlab with GPU development done in C++, using nVidia CUDA libraries. This approach was effective for rapid prototyping and proof of concept development. However, in future it would be useful to integrate the segmentation and annotation systems with widely used software, such as TrakEM2, and the other systems discussed in section 1.5.5, so that it can be of maximum benefit to the scientific community.

Approaches to image analysis based on receptive fields are inspired by research into the visual systems especially of mammals [44] and insects [92], in which visual interneurons have been shown to respond to bars, lines or edges. That area of vision research is under constant evaluation, and advances in it can lead to improved accuracy for segmentation and feature detection. Future avenues of research include identifying additional useful receptive field types and combining outputs from different receptive fields into a layered system for more accurate detection of cell membranes and other

organelles. Applying these techniques to ssTEM data offers for the future an improved understanding not only of visual systems but in turn also a further improvement of such computational techniques.

Chapter 6

Discussion

This research has made progress towards the goal of investigating functional units spanning multiple levels of biological organization in the brain. Linking information from different levels, or modalities of understanding in neuroscience is essential for building a more accurate picture of brain function, and brain disease. In Chapter 3, the BrainTrap database spearheaded this research into protein localization imaging and linked genetics and proteomics to anatomy in the fruit fly brain. In Chapter 4, network analysis techniques were developed and used to further link protein expression data with protein interaction networks, mouse gene expression patterns and connectivity networks. And in Chapter 5, a biologically inspired approach to EM image analysis was developed to assist in the task of connectivity data annotation. In the following sections I discuss future work that would continue and extend this research, and some areas for new research projects suggested by some of the work described in this thesis. Finally, in section 6.4, I summarize the main achievements and conclusions.

6.1 Protein Imaging

The BrainTrap website provides a valuable resource for research in the fruit fly brain, and provides insights into how biomolecular and anatomical modalities interact. The protein trap method used to image *Drosophila* brains under confocal microscopy in Chapter 3 has helped to bridge the gap between biomolecular and anatomical understanding. One obvious next step is to continue to image different proteins using techniques such as the protein trap method. As more proteins are imaged in the brain we will gain more insight into the interactions between proteins and the relationships between proteins and anatomy. In future, the set of 535 protein images in the BrainTrap

database could be extended with additional proteins from labs around the world [59], or the generation of further protein trap lines.

With greater coverage across the fruit fly genome, protein expression information will become even more valuable by adding more precision to protein interaction networks, and further linking biomolecular and anatomical modalities.

Protein trap lines already imaged are also useful for further experimental investigations into the trapped proteins, some of which are important for brain functions as outlined in Section 3.1.3. Affinity purification combined with mass spectrometry is one useful technique for discovering more about protein interactions with the protein trap lines [107]. For example, by using this method to target interactions in the brain, with synaptic protein trap lines, it would be possible to discover more details of synaptic protein interactions, and with this build up a picture of how the synapse works.

Attempts to image EYFP protein trap lines under EM as described in section 3.2 were not as successful as I would have liked but some progress was made and there are many options to pursue this line of investigation in future. Those options include perfecting a protocol to image EYFP in the *Drosophila* brain, and using different fixation conditions or staining techniques. Alternatively, the EYFP construct could be swapped, using additional genetic modification, for a protein more readily imaged at the EM level. If successful, this type of investigation could provide valuable information as it would further highlight potentially spurious protein-protein interactions, and would allow links between biomolecular, anatomical and connectivity levels of research that are not yet possible. Knowing which proteins localize to the synapse and how they interact *in vivo* is important information for computational models of synaptic plasticity, and for research into synaptopathies such as Alzheimer's disease and schizophrenia, for example.

6.2 Network Analysis

Databases containing biological information such as genetic sequences, protein interactions and evolutionary homology are vital tools for investigating the multiple levels of biological organization in the brain. The network analysis methods used and developed in Chapter 4 offer some techniques for improving the quality of data, such as protein interaction networks, and for gaining insights from the connections in the data, such as automatically segmenting the fruit fly brain based only on the protein expression profiles of different regions. Using evolutionary homology it was possible

to measure correlations between fly and mouse brain expression patterns and compare variability of ancient and recently evolved proteins in the fruit fly brain.

The amount of genetic and biological information available in online databases is overwhelming when compared with what was available 10 years ago, and we can expect this data to continue to both grow, and become more precise in the future. All the analysis techniques used in Chapter 4 will gain more statistical power when based on additional data. Therefore, in the future we can expect to be able to use the same analysis techniques with greater power to find further correlations between species and to identify genetically specific regions in the *Drosophila* brain with greater precision, for example.

Enhanced motif analysis, introduced in section 4.4.1, offers one way to further investigate connectivity data in conjunction with other data such as neuron type, or protein expression information. In future it is likely that more connectivity data will be available and this technique will be able to shed further light on the principles used to wire up the brain.

A full understanding of the interplay between the many levels of biological organization in the brain will require further development of these neuroinformatics tools to analyze the available data in a wholistic rather than reductionist way.

6.3 Tracing

Connectome research is currently struggling with the amount of manual annotation required to extract useful information from EM images. This annotation bottleneck limits connectome research to volumes so small that the modestly sized fruit fly brain is a huge undertaking [19]. The automatic tracing and alignment improvement systems described in Chapter 5 offer only partial relief from this bottleneck by allowing annotators to correct the imperfect automatically annotated images. For connectome research to scale beyond these limitations we will need to develop more accurate automatic annotation systems. The requirement for accuracy in these systems is very high; with image series reaching up to hundreds or even thousands of consecutive images, even one mistake per image will completely obfuscate the resulting connectivity network.

In the immediate future, the segmentation, alignment improvement and annotation software developed here will be integrated with existing software systems, such as TrakEM2, to provide maximum benefit to the scientific community. This type of project requires scientists, annotators and software developers working together to cre-

ate more usable and scalable systems that can be used by many different labs.

The receptive fields approach described in Chapter 5, inspired by biological visual systems, offers many avenues for future research. For example, taking further insight from the area of biological vision, future research could identify additional useful receptive field types, and combine outputs from different receptive fields into a layered system, modeled on the layered vision systems observed in mammalian brains. Mitochondria and synapse feature detection could also be extended to recognize vesicles, as discussed in section 5.8, and other organelles such as ribosomes, cell nuclei and the axon initial segment. All of these elements are potentially important for fully understanding the connectome data, so measuring them automatically will provide valuable information and save time required for researchers to interpret EM data.

6.4 Conclusion

Understanding the brain is an immense challenge. Even the fruit fly with only 100,000 neurons in its brain demonstrates complicated behavior which cannot be replicated by the use of computational modeling or artificial intelligence. Some basic functions have been studied and replicated in computational or robotic models, but the full range of behavior and the speed and accuracy at which the fruit fly can perform them is far beyond our current abilities to replicate *in silico*. Further understanding the fruit fly brain gives us some idea, even at this basic level, of what biological intelligence is. Additionally, because of our common evolutionary past and genetic similarities, it provides us a tool with which to understand mammalian brain function at a biological level, which will aid future research into brain diseases and provide a stepping stone for furthering our understanding of human intelligence.

Anatomical protein localization is one area where further information is needed for tying together the genetic, anatomical and connectivity levels. The new BrainTrap database spearheads this research, with all information gathered publicly available and presented in a format which is easy to use and search. Using an agreed upon ontology to mark up expression patterns, BrainTrap demonstrates how public databases can be linked to relevant resources and standards. In future, the BrainTrap website will be maintained as a resource for the community, and any compatible data that is made available will also be included on the website.

The BrainTrap data provides one example of information that can join multiple levels of abstraction used for describing and understanding the brain. Biomolecular and

anatomical levels come together in this data and this can help us to further understand both levels.

The annotation data in BrainTrap was used to improve protein-protein interaction data for the fly brain. A total of 149 interactions can be ruled out based on only 535 protein trap lines. With additional protein expression pattern information the number of ruled out interactions would grow exponentially. This information combined with other protein research will greatly improve our currently limited understanding of protein-protein interactions that occur in the brain.

The MultiBrain visualization feature demonstrates further the potential for protein expression data. Once mapped to a common template, protein expression patterns can be directly compared in a single image. Future work on registered brain volumes could also achieve automatic annotation of 3D expression patterns. The BrainTrap data would provide an ideal test set of data for such a project.

By using a selection of protein expression patterns mapped to a template, it was also possible to create informative segmentations of *Drosophila* brain areas. With further work in this area it may be possible to produce a more detailed brain atlas and anatomical ontology based on brain region similarities at the molecular level. It will also be possible to link detailed protein expression profiles to each area of the brain.

In addition to data sharing and visualization, the BrainTrap database contains annotations useful for combining this data with other sources of information. This was demonstrated in section 4.3.1 with data from the Allen Brain Atlas. Even with the limited number of protein expression patterns available, we discovered correlations between brain areas in fly and mouse. In section 4.3.2, using homology data along with the BrainTrap annotations, it was also possible to investigate the evolution of brain proteins. Further analysis and more comprehensive databases of protein expression will in future provide an even more detailed understanding of these relationships and our common evolutionary history.

Linking protein expression data to the connectivity level has potential to unlock a much more detailed picture of the organization of the brain. Tools such as the enhanced motif method introduced in Section 4.4.1 demonstrate how this can be done. Currently however, there is very little connectivity data available and this cannot yet be investigated for large connectivity networks. With much effort currently going into EM reconstruction techniques this may not be the case in the future.

Connectivity data is difficult to acquire and time consuming to annotate. Computer algorithms are still unable to come close to gold standard accuracy levels which hu-

mans routinely achieve for ssTEM images. I have contributed towards the automatic annotation of high resolution connectivity data by implementing a segmentation and alignment improvement system that provides an improvement over existing software. Based on aspects of biological vision systems, the density2d ridge detection system has potential for further improvement by integrating further aspects of biological vision into the system. Combining receptive field and ridge detection approaches with other methods, such as random forest classifiers or the watershed algorithm, provides additional areas for further research.

GPU technology has allowed these principles to be extended into full 3D image analysis for neuron tracing in both EM and light microscopy data, as demonstrated in section 5.4. The increased computing power available in GPU computing is encouraging for these computationally intensive methods. Microscopy techniques and EM image quality and resolution are also improving and allowing very large volumes of data to be imaged at this high resolution. Manual annotation of this data will not scale at the same rate that data acquisition will. Fully automatic tracing should therefore be the goal for future research in connectomics.

In this thesis I have discussed several levels at which we can understand the brain: molecular biology, anatomy, and connectivity. Research has traditionally focused on one level at a time. Linking these levels of knowledge is essential to further our understanding of the brain and discover hidden interactions between modalities, as demonstrated by this research.

Bibliography

[1] Mark D. Adams, Susan E. Celniker, Robert A. Holt, Cheryl A. Evans, Jeannine D. Gocayne, Peter G. Amanatides, Steven E. Scherer, Peter W. Li, Roger A. Hoskins, Richard F. Galle, Reed A. George, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, Scott N. Henderson, Granger G. Sutton, Jennifer R. Wortman, Mark D. Yandell, Qing Zhang, Lin X. Chen, Rhonda C. Brandon, Yu-Hui C. Rogers, Robert G. Blazej, Mark Champe, Barret D. Pfeiffer, Kenneth H. Wan, Clare Doyle, Evan G. Baxter, Gregg Helt, Catherine R. Nelson, George L. Gabor, Miklos, Josep F. Abril, Anna Agbayani, Hui-Jin An, Cynthia Andrews-Pfannkoch, Danita Baldwin, Richard M. Ballew, Anand Basu, James Baxendale, Leyla Bayraktaroglu, Ellen M. Beasley, Karen Y. Beeson, P. V. Benos, Benjamin P. Berman, Deepali Bhandari, Slava Bolshakov, Dana Borkova, Michael R. Botchan, John Bouck, Peter Brokstein, Phillipe Brottier, Kenneth C. Burtis, Dana A. Busam, Heather Butler, Edouard Cadieu, Angela Center, Ishwar Chandra, J. Michael Cherry, Simon Cawley, Carl Dahlke, Lionel B. Davenport, Peter Davies, Beatriz de Pablos, Arthur Delcher, Zuoming Deng, Anne D. Mays, Ian Dew, Suzanne M. Dietz, Kristina Dodson, Lisa E. Doup, Michael Downes, Shannon Dugan-Rocha, Boris C. Dunkov, Patrick Dunn, Kenneth J. Durbin, Carlos C. Evangelista, Concepcion Ferraz, Steven Ferriera, Wolfgang Fleischmann, Carl Fosler, Andrei E. Gabrielian, Neha S. Garg, William M. Gelbart, Ken Glasser, Anna Glodek, Fangcheng Gong, J. Harley Gorrell, Zhiping Gu, Ping Guan, Michael Harris, Nomi L. Harris, Damon Harvey, Thomas J. Heiman, Judith R. Hernandez, Jarrett Houck, Damon Hostin, Kathryn A. Houston, Timothy J. Howland, Ming-Hui Wei, Chinyere Ibegwam, Mena Jalali, Francis Kalush, Gary H. Karpen, Zhaoxi Ke, James A. Kennison, Karen A. Ketchum, Bruce E. Kimmel, Chinnappa D. Kodira, Cheryl Kraft, Saul Kravitz, David Kulp, Zhongwu Lai, Paul Lasko, Yiding Lei, Alexander A. Levitsky, Jiayin Li, Zhenya Li, Yong Liang, Xiaoying Lin, Xiangjun Liu, Bettina Mattei, Tina C. McIntosh, Michael P. McLeod, Duncan McPherson, Gennady Merkulov, Natalia V. Milshina, Clark Mobarry, Joe Morris, Ali Moshrefi, Stephen M. Mount, Mee Moy, Brian Murphy, Lee Murphy, Donna M. Muzny, David L. Nelson, David R. Nelson, Keith A. Nelson, Katherine Nixon, Deborah R. Nusskern, Joanne M. Pacleb, Michael Palazzolo, Gjange S. Pittman, Sue Pan, John Pollard, Vinita Puri, Martin G. Reese, Knut Reinert, Karin Remington, Robert D. C. Saunders, Frederick Scheeler, Hua Shen, Bixiang C. Shue, Inga Sidén-Kiamos, Michael Simpson, Marian P. Skupski, Tom Smith, Eugene Spier, Allan C. Spradling, Mark Stapleton, Renee Strong, Eric Sun, Robert Svirskas, Cyndee Tector, Russell Turner, Eli Venter, Aihui H. Wang, Xin Wang,

Zhen-Yuan Wang, David A. Wassarman, George M. Weinstock, Jean Weissenbach, Sherita M. Williams, Trevor Woodage, Kim C. Worley, David Wu, Song Yang, Q. Alison Yao, Jane Ye, Ru-Fang Yeh, Jayshree S. Zaveri, Ming Zhan, Guangren Zhang, Qi Zhao, Liansheng Zheng, Xiangqun H. Zheng, Fei N. Zhong, Wenyan Zhong, Xiaojun Zhou, Shiaoping Zhu, Xiaohong Zhu, Hamilton O. Smith, Richard A. Gibbs, Eugene W. Myers, Gerald M. Rubin, and J. Craig Venter. The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, March 2000.

[2] Ryan Andersen, Yimei Li, Mary Resseguie, and Jay E. Brenman. Calcium/Calmodulin-Dependent Protein Kinase II Alters Structural Plasticity and Cytoskeletal Dynamics in Drosophila. *J. Neurosci.*, 25(39):8878–8888, September 2005.

[3] James R. Anderson, Bryan W. Jones, Jia-Hui Yang, Marguerite V. Shaw, Carl B. Watt, Pavel Koshevoy, Joel Spaltenstein, Elizabeth Jurrus, Kannan Uv, Ross T. Whitaker, David Mastronarde, Tolga Tasdizen, and Robert E. Marc. A Computational Framework for Ultrastructural Mapping of Neural Circuitry. *PLoS Biol*, 7(3):e1000074+, March 2009.

[4] Alessandra Angelucci, Jonathan B. Levitt, and Jennifer S. Lund. Anatomical origins of the classical receptive field and modulatory surround field of single neurons in macaque visual cortical area V1. *Progress in brain research*, 136:373–388, 2002.

[5] J. Douglas Armstrong and Jano I. van Hemert. Towards a virtual fly brain. *Physical and Engineering Sciences*, 367(1896):2387–2397, June 2009.

[6] Hugo J. Bellen, Robert W. Levis, Guochun Liao, Yuchun He, Joseph W. Carlson, Garson Tsang, Martha Evans-Holm, P. Robin Hiesinger, Karen L. Schulze, Gerald M. Rubin, Roger A. Hoskins, and Allan C. Spradling. The BDGP Gene Disruption Project: Single Transposon Insertions Associated With 40% of Drosophila Genes. *Genetics*, 167(2):761–781, June 2004.

[7] Kelly Beumer, Heinrich J. G. Matthies, Amber Bradshaw, and Kendal Broadie. Integrins regulate DLG/FAS2 via a CaM kinase II-dependent pathway to mediate synapse elaboration and stabilization during postembryonic development. *Development*, 129(14):3381–3391, July 2002.

[8] Davi D. Bock, Wei-Chung A. Lee, Aaron M. Kerlin, Mark L. Andermann, Greg Hood, Arthur W. Wetzel, Sergey Yurgenson, Edward R. Soucy, Hyon S. Kim, and R. Clay Reid. Network anatomy and in vivo physiology of visual cortical neurons. *Nature*, 471(7337):177–182, March 2011.

[9] Jason W. Bohland, Hemant Bokil, Sayan D. Pathak, Chang-Kyu Lee, Lydia Ng, Christopher Lau, Chihchau Kuan, Michael Hawrylycz, and Partha P. Mitra. Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods*, 50(2):105–112, February 2010.

[10] Kevin L. Briggman and Winfried Denk. Towards neural circuit reconstruction with volume electron microscopy techniques. *Current Opinion in Neurobiology*, 16(5):562–570, October 2006.

[11] Kevin L. Briggman, Moritz Helmstaedter, and Winfried Denk. Wiring specificity in the direction-selectivity circuit of the retina. *Nature*, 471(7337):183–188, March 2011.

[12] Sebastian Cachero, Aaron D. Ostrovsky, Jai Y. Yu, Barry J. Dickson, and Gregory S. Jefferis. Sexual dimorphism in the fly brain. *Current biology : CB*, 20(18):1589–1601, September 2010.

[13] Matteo Carandini, Jonathan B. Demb, Valerio Mante, David J. Tolhurst, Yang Dan, Bruno A. Olshausen, Jack L. Gallant, and Nicole C. Rust. Do We Know What the Early Visual System Does? *The Journal of Neuroscience*, 25(46):10577–10597, November 2005.

[14] Albert Cardona, Stephan Saalfeld, Ignacio Arganda, Wayne Pereanu, Johannes Schindelin, and Volker Hartenstein. Identifying Neuronal Lineages of Drosophila by Sequence Analysis of Axon Tracts. *The Journal of Neuroscience*, 30(22):7538–7553, June 2010.

[15] Albert Cardona, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Tomancak, and Volker Hartenstein. An Integrated Micro- and Macroarchitectural Analysis of the Drosophila Brain by Computer-Assisted Serial Section Electron Microscopy. *PLoS Biol*, 8(10):e1000502+, October 2010.

[16] Alicia M. Celotto and Michael J. Palladino. Drosophila: a "model" model system to study neurodegeneration. *Molecular interventions*, 5(5):292–303, October 2005.

[17] Francis C. Chee, Amritpal Mudher, Matthew F. Cuttle, Tracey A. Newman, Daniel MacKay, Simon Lovestone, and David Shepherd. Over-expression of tau results in defective synaptic transmission in Drosophila neuromuscular junctions. *Neurobiology of Disease*, 20(3):918–928, December 2005.

[18] Ann-Shyn Chiang, Chih-Yung Lin, Chao-Chun Chuang, Hsiu-Ming Chang, Chang-Huain Hsieh, Chang-Wei Yeh, Chi-Tin Shih, Jian-Jheng Wu, Guo-Tzau Wang, Yung-Chang Chen, Cheng-Chi Wu, Guan-Yu Chen, Yu-Tai Ching, Ping-Chang Lee, Chih-Yang Lin, Hui-Hao Lin, Chia-Chou Wu, Hao-Wei Hsu, Yun-Ann Huang, Jing-Yi Chen, Hsin-Jung Chiang, Chun-Fang Lu, Ru-Fen Ni, Chao-Yuan Yeh, and Jenn-Kang Hwang. Three-Dimensional Reconstruction of Brain-wide Wiring Networks in Drosophila at Single-Cell Resolution. *Current Biology*, 21(1):1–11, January 2011.

[19] Dmitri B. Chklovskii, Shiv Vitaladevuni, and Louis K. Scheffer. Semi-automated reconstruction of neural circuits using electron microscopy. *Current opinion in neurobiology*, 20(5):667–675, October 2010.

[20] Sommer Christoph, Straehle Christoph, Köthe Ullrich, and A. Hamprecht Fred. ilastik: Interactive Learning and Segmentation Toolkit. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, March 2011.

[21] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[22] Nuno M. da Costa, David Fürsinger, and Kevan A. C. Martin. The Synaptic Organization of the Claustral Projection to the Cat's Visual Cortex. *The Journal of Neuroscience*, 30(39):13166–13170, September 2010.

[23] J. Dauguet, D. Bock, R. C. Reid, and S. K. Warfield. Alignment of large image series using cubic B-splines tessellation: application to transmission electron microscopy data. *Medical image computing and computer-assisted intervention : MICCAI*, 10(Pt 2):710–717, 2007.

[24] Ronald L. Davis. OLFACTORY MEMORY FORMATION IN DROSOPHILA: From Molecular to Systems Neuroscience. *Annual Review of Neuroscience*, 28(1):275–302, 2005.

[25] Michael Eisenstein. Neural circuits: Putting neurons on the map. *Nature*, 461(7267):1149–1152, October 2009.

[26] Richard D. Emes, Andrew J. Pocklington, Christopher N. Anderson, Alex Bayes, Mark O. Collins, Catherine A. Vickers, Mike D. Croning, Bilal R. Malik, Jyoti S. Choudhary, J. Douglas Armstrong, and Seth G. Grant. Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nature neuroscience*, 11(7):799–806, July 2008.

[27] Lina Enell, Yasutaka Hamasaka, Agata Kolodziejczyk, and Dick R. Nässel. Gamma-Aminobutyric acid (GABA) signaling components in Drosophila: Immunocytochemical localization of GABAB receptors in relation to the GABAA receptor subunit RDL and a vesicular GABA transporter. *J. Comp. Neurol.*, 505(1):18–31, 2007.

[28] John C. Fiala. Reconstruct: a free editor for serial section microscopy. *Journal of microscopy*, 218(Pt 1):52–61, April 2005.

[29] S. Fields and R. Sternglanz. The two-hybrid system: an assay for protein-protein interactions. *Trends in genetics : TIG*, 10(8):286–292, August 1994.

[30] C. Fink. Molecular mechanisms of CaMKII activation in neuronal plasticity. *Current Opinion in Neurobiology*, 12(3):293–299, June 2002.

[31] David Fitzpatrick. Seeing beyond the receptive field in primary visual cortex. *Current Opinion in Neurobiology*, 10(4):438–443, August 2000.

[32] Mark E. Fortini, Marian P. Skupski, Mark S. Boguski, and Iswar K. Hariharan. A Survey of Human Disease Gene Counterparts in the Drosophila Genome. *J. Cell Biol.*, 150(2):23–30, July 2000.

[33] Katsuo Furukubo-Tokunaga. *Modeling schizophrenia in flies*, volume 179, pages 107–115. Elsevier, 2009.

[34] Gary Grumbling, Victor Strelets, and The FlyBase Consortium. FlyBase: anatomical data, images and queries. *Nucl. Acids Res.*, 34(suppl_1):D484–488, January 2006.

[35] Udo Hacker, Sverker Nystedt, Mojgan P. Barmchi, Carsten Horn, and Ernst A. Wimmer. piggyBac-based insertional mutagenesis in the presence of stably integrated P elements in Drosophila. *PNAS*, 100(13):7720–7725, June 2003.

[36] Dafni Hadjieconomou, Shay Rotkopf, Cyrille Alexandre, Donald M. Bell, Barry J. Dickson, and Iris Salecker. Flybow: genetic multicolor cell labeling for neural circuit analysis in Drosophila melanogaster. *Nature Methods*, 8(3):260–266, February 2011.

[37] Stefanie Hampel, Phuong Chung, Claire E. McKellar, Donald Hall, Loren L. Looger, and Julie H. Simpson. Drosophila Brainbow: a recombinase-based fluorescence labeling technique to subdivide neural expression patterns. *Nature Methods*, 8(3):253–259, February 2011.

[38] H. Keffer Hartline, Henry G. Wagner, and Floyd Ratliff. Inhibition in the eye of *Limulus*. *The Journal of general physiology*, 39(5):651–673, May 1956.

[39] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, new edition edition, June 1949.

[40] Gena Heidary and Mark E. Fortini. Identification and characterization of the Drosophila tau homolog. *Mechanisms of Development*, 108(1-2):171–178, October 2001.

[41] Moritz Helmstaedter, Kevin L. Briggman, and Winfried Denk. High-accuracy neurite reconstruction for high-throughput neuroanatomy. *Nature Neuroscience*, 14(8):1081–1088, July 2011.

[42] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, August 1952.

[43] Carsten Horn, Nils Offen, Sverker Nystedt, Udo Hacker, and Ernst A. Wimmer. piggyBac-Based Insertional Mutagenesis and Enhancer Detection as a Tool for Functional Insect Genomics. *Genetics*, 163(2):647–661, February 2003.

[44] David H. Hubel and Torsten N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148:574–591, October 1959.

[45] Jennifer M. Ichida, Lars Schwabe, Paul C. Bressloff, and Alessandra Angelucci. Response Facilitation From the "Suppressive" Receptive Field Surround of Macaque V1 Neurons. *J Neurophysiol*, 98(4):2168–2181, October 2007.

[46] Konstantin G. Iliadi, Aaron Avivi, Natalia N. Iliadi, David Knight, Abraham B. Korol, Eviatar Nevo, Paul Taylor, Michael F. Moran, Nikolai G. Kamyshev, and Gabrielle L. Boulianne. nemy encodes a cytochrome b561 that is required for Drosophila learning and memory. *Proceedings of the National Academy of Sciences*, 105(50):19986–19991, December 2008.

[47] George R. Jackson, Martina Wiedau-Pazos, Tzu-Kang Sang, Naveed Wagle, Carlos A. Brown, Sasan Massachi, and Daniel H. Geschwind. Human Wild-Type Tau Interacts with wingless Pathway Components and Produces Neurofibrillary Pathology in Drosophila. *Neuron*, 34(4):509–519, May 2002.

[48] Viren Jain, H. Sebastian Seung, and Srinivas C. Turaga. Machines that learn to segment images: a crucial technology for connectomics. *Current opinion in neurobiology*, 20(5):653–666, October 2010.

[49] Gregory S. Jefferis, Christopher J. Potter, Alexander M. Chan, Elizabeth C. Marin, Torsten Rohlfing, Calvin R. Maurer, and Liqun Luo. Comprehensive maps of Drosophila higher olfactory centers: spatially segregated fruit and pheromone representation. *Cell*, 128(6):1187–1203, March 2007.

[50] Astrid Jeibmann and Werner Paulus. Drosophila melanogaster as a model organism of brain diseases. *International journal of molecular sciences*, 10(2):407–440, February 2009.

[51] Arnim Jenett, Johannes E. Schindelin, and Martin Heisenberg. The Virtual Insect Brain Protocol: creating and comparing standardized neuroanatomy. *BMC Bioinformatics*, 7:544+, December 2006.

[52] Xiaoyi Jiang, H. Bunke, K. Abegglen, and A. Kandel. Curve morphing by weighted mean of strings. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 192–195. IEEE, 2002.

[53] Lily Kahsai and Åsa M. E. Winther. Chemical neuroanatomy of the Drosophila central complex: Distribution of multiple neuropeptides in relation to neurotransmitters. *J. Comp. Neurol.*, 519(2):290–315, 2011.

[54] Nikolai Kamyshev, Konstantin Iliadi, Julia Bragina, Elena Kamysheva, Elena Tokmatcheva, Thomas Preat, and Elena S. Popova. Novel memory mutants in Drosophila: Behavioral characteristics of the mutant nemyP153. *BMC Neuroscience*, 3(1):9+, 2002.

[55] Jan Karbowski, Gary Schindelman, Christopher Cronin, Adeline Seah, and Paul Sternberg. Systems level circuit model of C. elegans undulatory locomotion: mathematical modeling and molecular genetics. *Journal of Computational Neuroscience*, 24(3):253–276, June 2008.

[56] Yael Katz, Vilas Menon, Daniel A. Nicholson, Yuri Geinisman, William L. Kath, and Nelson Spruston. Synapse Distribution Suggests a Two-Stage Model of Dendritic Integration in CA1 Pyramidal Neurons. *Neuron*, 63(2):171–177, July 2009.

[57] Verena Kaynig, Thomas Fuchs, and Joachim M. Buhmann. Neuron geometry extraction by perceptual grouping in ssTEM images. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2902–2909, June 2010.

[58] Krystyna Keleman, Sebastian Kruttner, Mattias Alenius, and Barry J. Dickson. Function of the Drosophila CPEB protein Orb2 in long-term courtship memory. *Nature Neuroscience*, 10(12):1587–1593, October 2007.

[59] R. J. Kelso, M. Buszczak, A. T. Quiñones, C. Castiblanco, S. Mazzalupo, and L. Cooley. Flytrap, a database documenting a GFP protein-trap insertion screen in Drosophila melanogaster. *Nucleic Acids Res*, 32(Database issue), January 2004.

[60] Ken-ichi Kimura, Tomoaki Hachiya, Masayuki Koganezawa, Tatsunori Tazawa, and Daisuke Yamamoto. Fruitless and Doublesex Coordinate to Generate Male-Specific Neurons that Can Initiate Courtship. *Neuron*, 59(5):759–769, September 2008.

[61] Graham Knott, Herschel Marchman, David Wall, and Ben Lich. Serial section scanning electron microscopy of adult brain tissue using focused ion beam milling. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 28(12):2959–2964, March 2008.

[62] Seymour Knowles-Barley, Nancy J. Butcher, Ian A. Meinertzhagen, and J. Douglas Armstrong. Biologically inspired EM image alignment and neural reconstruction. *Bioinformatics*, 27(16):2216–2223, August 2011.

[63] Seymour Knowles-Barley, Mark Longair, and J. Douglas Armstrong. BrainTrap: a database of 3D protein expression patterns in the Drosophila brain. *Database*, 2010(0):baq005+, March 2010.

[64] Young H. Koh, Evgenya Popova, Ulrich Thomas, Leslie C. Griffith, and Vivian Budnik. Regulation of DLG Localization at Synapses by CaMKII-Dependent Phosphorylation. *Cell*, 98(3):353–363, August 1999.

[65] Kenneth S. Kosik. Beyond phrenology, at Last. *Nat Rev Neurosci*, 4(3):234–239, March 2003.

[66] James R. Kremer, David N. Mastronarde, and J. Richard McIntosh. Computer Visualization of Three-Dimensional Image Data Using IMOD. *Journal of Structural Biology*, 116(1):71–76, January 1996.

[67] Doris Kretzschmar. Neurodegenerative mutants in Drosophila: a means to identify genes and mechanisms involved in human diseases? *Invertebrate Neuroscience*, 5(3):97–109, November 2005.

[68] Lars V. Kristiansen and Michael Hortsch. Fasciclin II: The NCAM Ortholog in <i>Drosophila melanogaster</i> Structure and Function of the Neural Cell Adhesion Molecule NCAM. volume 663 of *Advances in Experimental*

Medicine and Biology, chapter 24, pages 387–401. Springer New York, New York, NY, 2010.

[69] Mitsuhiko Kurusu, Takeshi Awasaki, Liria M. Masuda-Nakagawa, Hiroshi Kawauchi, Kei Ito, and Katsuo Furukubo-Tokunaga. Embryonic and larval development of the Drosophila mushroom bodies: concentric layer subdivisions and the role of fasciclin II. *Development*, 129(2):409–419, January 2002.

[70] Ed S. Lein, Michael J. Hawrylycz, Nancy Ao, Mikael Ayres, Amy Bensinger, Amy Bernard, Andrew F. Boe, Mark S. Boguski, Kevin S. Brockway, Emi J. Byrnes, Lin Chen, Li Chen, Tsuey-Ming Chen, Mei C. Chin, Jimmy Chong, Brian E. Crook, Aneta Czaplinska, Chinh N. Dang, Suvro Datta, Nick R. Dee, Aimee L. Desaki, Tsega Desta, Ellen Diep, Tim A. Dolbeare, Matthew J. Donelan, Hong-Wei Dong, Jennifer G. Dougherty, Ben J. Duncan, Amanda J. Ebbert, Gregor Eichele, Lili K. Estin, Casey Faber, Benjamin A. Facer, Rick Fields, Shanna R. Fischer, Tim P. Fliss, Cliff Frensley, Sabrina N. Gates, Katie J. Glattfelder, Kevin R. Halverson, Matthew R. Hart, John G. Hohmann, Maureen P. Howell, Darren P. Jeung, Rebecca A. Johnson, Patrick T. Karr, Reena Kawal, Jolene M. Kidney, Rachel H. Knapik, Chihchau L. Kuan, James H. Lake, Annabel R. Laramee, Kirk D. Larsen, Christopher Lau, Tracy A. Lemon, Agnes J. Liang, Ying Liu, Lon T. Luong, Jesse Michaels, Judith J. Morgan, Rebecca J. Morgan, Marty T. Mortrud, Nerick F. Mosqueda, Lydia L. Ng, Randy Ng, Geralyn J. Orta, Caroline C. Overly, Tu H. Pak, Sheana E. Parry, Sayan D. Pathak, Owen C. Pearson, Ralph B. Puchalski, Zackery L. Riley, Hannah R. Rockett, Stephen A. Rowland, Joshua J. Royall, Marcos J. Ruiz, Nadia R. Sarno, Katherine Schaffnit, Nadiya V. Shapovalova, Taz Sivisay, Clifford R. Slaughterbeck, Simon C. Smith, Kimberly A. Smith, Bryan I. Smith, Andy J. Sodt, Nick N. Stewart, Kenda-Ruth Stumpf, Susan M. Sunkin, Madhavi Sutram, Angelene Tam, Carey D. Teemer, Christina Thaller, Carol L. Thompson, Lee R. Varnam, Axel Visel, Ray M. Whitlock, Paul E. Wohnoutka, Crissa K. Wolkey, Victoria Y. Wong, Matthew Wood, Murat B. Yaylaoglu, Rob C. Young, Brian L. Youngstrom, Xu F. Yuan, Bin Zhang, Theresa A. Zwingman, and Allan R. Jones. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, January 2007.

[71] Florian Leiss, Claudia Groh, Nancy J. Butcher, Ian A. Meinertzhagen, and Gaia Tavosanis. Synaptic organization in the adult Drosophila mushroom body calyx. *The Journal of Comparative Neurology*, 517(6):808–824, 2009.

[72] Jianli Li, Yue Wang, Shu-Ling L. Chiu, and Hollis T. Cline. Membrane targeted horseradish peroxidase as a marker for correlative fluorescence and electron microscopy studies. *Frontiers in neural circuits*, 4, 2010.

[73] Jeff W. Lichtman, Jean Livet, and Joshua R. Sanes. A technicolour approach to the connectome. *Nature reviews. Neuroscience*, 9(6):417–422, June 2008.

[74] Jeff W. Lichtman and Joshua R. Sanes. Ome sweet ome: what can the genome tell us about the connectome? *Current opinion in neurobiology*, 18(3):346–353, June 2008.

[75] Susana Q. Lima and Gero Miesenböck. Remote control of behavior through genetically targeted photostimulation of neurons. *Cell*, 121(1):141–152, April 2005.

[76] Jean Livet, Tamily A. Weissman, Hyuno Kang, Ryan W. Draft, Ju Lu, Robyn A. Bennis, Joshua R. Sanes, and Jeff W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, November 2007.

[77] Ju Lu, John C. Fiala, and Jeff W. Lichtman. Semi-automated reconstruction of neural processes from large numbers of fluorescence images. *PloS one*, 4(5):e5655+, May 2009.

[78] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-Based Segmentation of EM Image Stacks with Learned Shape Features. Technical report, EPFL, 2010.

[79] E. R. Macagno, C. Levinthal, and I. Sobel. Three-Dimensional Computer Reconstruction of Neurons and Neuronal Assemblies. *Annual Review of Biophysics and Bioengineering*, 8(1):323–351, 1979.

[80] Trudy F. C. Mackay and Robert R. H. Anholt. Of Flies and Man: Drosophila as a Model for Human Complex Traits. *Annual Review of Genomics and Human Genetics*, 7(1):339–367, 2006.

[81] Dieter Maier, Angelika Hausser, Anja C. Nagel, Gisela Link, Sabrina J. Kugler, Irmgard Wech, Klaus Pfizenmaier, and Anette Preiss. Drosophila protein kinase D is broadly expressed and a fraction localizes to the Golgi compartment. *Gene Expression Patterns*, 6(8):849–856, October 2006.

[82] Hari Manev and Svetlana Dzitoyeva. *GABA-B Receptors in Drosophila*, volume 58, pages 453–464. Elsevier, 2010.

[83] Tomoko Mastushita-Sakai, Erica White-Grindley, Jessica Samuelson, Chris Seidel, and Kausik Si. Drosophila Orb2 targets genes involved in neuronal growth, synapse formation, and protein turnover. *Proceedings of the National Academy of Sciences*, 107(26):11987–11992, June 2010.

[84] Ian A. Meinertzhagen and S. D. O'Neil. Synaptic organization of columnar elements in the lamina of the wild type in Drosophila melanogaster. *The Journal of Comparative Neurology*, 305(2):232–263, March 1991.

[85] Angel Merchán-Pérez, José-Rodrigo R. Rodriguez, Lidia Alonso-Nanclares, Andreas Schertel, and Javier Defelipe. Counting Synapses Using FIB/SEM Microscopy: A True Revolution for Ultrastructural Volume Reconstruction. *Frontiers in neuroanatomy*, 3, 2009.

[86] Gero Miesenböck and Ioannis G. Kevrekidis. Optical imaging and control of genetically designated neurons in functioning circuits. *Annual review of neuroscience*, 28:533–563, 2005.

[87] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, October 2002.

[88] Yuriy Mishchenko. Automation of 3D reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs. *Journal of Neuroscience Methods*, 176(2):276–289, January 2009.

[89] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[90] Lydia Ng, Chris Lau, Rob Young, Sayan Pathak, Leonard Kuan, Andrew Sodt, Madhavi Sutram, Chang K. Lee, Chinh Dang, and Michael Hawrylycz. NeuroBlast: a 3D spatial homology search tool for gene expression. *BMC Neuroscience*, 8(Suppl 2):P11+, 2007.

[91] J. Nickolls and W. J. Dally. The GPU Computing Era. *Micro, IEEE*, 30(2):56–69, March 2010.

[92] David O'Carroll. Feature-detecting neurons in dragonflies. *Nature*, 362(6420):541–543, April 1993.

[93] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.

[94] Bruno A. Olshausen and David J. Field. How Close Are We to Understanding V1? *Neural Computation*, 17(8):1665–1699, August 2005.

[95] Hideo Otsuna and Kei Ito. Systematic analysis of the visual projection neurons of Drosophila melanogaster. I. Lobula-specific pathways. *The Journal of comparative neurology*, 497(6):928–958, August 2006.

[96] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. GPU Computing. *Proceedings of the IEEE*, 96(5):879–899, May 2008.

[97] Hirofumi Ozeki, Ian M. Finn, Evan S. Schaffer, Kenneth D. Miller, and David Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–592, May 2009.

[98] Hirofumi Ozeki, Osamu Sadakane, Takafumi Akasaki, Tomoyuki Naito, Satoshi Shimegi, and Hiromichi Sato. Relationship between Excitation and Inhibition Underlying Size Tuning and Contextual Response Modulation in the Cat Primary Visual Cortex. *The Journal of Neuroscience*, 24(6):1428–1438, February 2004.

[99] Ganna Palagina, Ulf T. Eysel, and Dirk Jancke. Strengthening of lateral activation in adult rat visual cortex after retinal lesions captured with voltage-sensitive dye imaging in vivo. *Proceedings of the National Academy of Sciences*, 106(21):8743–8747, May 2009.

[100] Demian Park, Melissa J. Coleman, James J. L. Hodge, Vivian Budnik, and Leslie C. Griffith. Regulation of neuronal excitability in Drosophila by constitutively active CaMKII. *J. Neurobiol.*, 52(1):24–42, 2002.

[101] Dorit Parnas, Pejmun A. Haghighi, Richard D. Fetter, Sang W. Kim, and Corey S. Goodman. Regulation of Postsynaptic Structure and Protein Localization by the Rho-Type Guanine Nucleotide Exchange Factor dPix. *Neuron*, 32(3):415–424, November 2001.

[102] Jodi R. Parrish, Keith D. Gulyas, and Russell L. Finley. Yeast two-hybrid contributions to interactome mapping. *Current opinion in biotechnology*, 17(4):387–393, August 2006.

[103] Hanchuan Peng, Phuong Chung, Fuhui Long, Lei Qu, Arnim Jenett, Andrew M. Seeds, Eugene W. Myers, and Julie H. Simpson. BrainAligner: 3D registration atlases of Drosophila brains. *Nature Methods*, 8(6):493–498, May 2011.

[104] Hanchuan Peng, Fuhui Long, and Gene Myers. Automatic 3D neuron tracing using all-path pruning. *Bioinformatics*, 27(13):i239–i247, July 2011.

[105] Hanchuan Peng, Zongcai Ruan, Fuhui Long, Julie H. Simpson, and Eugene W. Myers. V3D enables real-time 3D visualization and quantitative analysis of large-scale biological image data sets. *Nature biotechnology*, 28(4):348–353, April 2010.

[106] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[107] Johanna S. Rees, Nick Lowe, Irina M. Armean, John Roote, Glynnis Johnson, Emma Drummond, Helen Spriggs, Edward Ryder, Steven Russell, Daniel St. Johnston, and Kathryn S. Lilley. In vivo analysis of proteomes and interactomes using parallel affinity capture (iPAC) coupled to mass spectrometry. *Molecular & Cellular Proteomics*, March 2011.

[108] Markus Reigl, Uri Alon, and Dmitri Chklovskii. Search for computational modules in the C. elegans brain. *BMC Biology*, 2(1):25+, 2004.

[109] L. T. Reiter, L. Potocki, S. Chien, M. Gribskov, and E. Bier. A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. *Genome research*, 11(6):1114–1125, June 2001.

[110] Torsten Rohlfing and Calvin R. Maurer. Nonrigid image registration in shared-memory multiprocessor environments with application to brains, breasts, and bees. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, 7(1):16–25, March 2003.

[111] Glenn D. Rosen, Alexander G. Williams, Anthony J. Capra, Michael T. Connolly, Brian Cruz, Lu Lu, David C. Airey, Anand Kulkarni, and Robert W. Williams. The Mouse Brain Library @ www.mbl.org. In *14th International Mouse Genome Meeting*, Narita, Japan, November 2000.

[112] E. Ryder, H. Spriggs, E. Drummond, D. St Johnston, and S. Russell. The Flannotator–a gene and protein expression annotation tool for Drosophila melanogaster. *Bioinformatics*, 25(4):548–549, February 2009.

[113] E. Ryder, H. Spriggs, G. Johnson, E. Drummond, J. Drummond, J. Webster, J. Roote, N. Lowe, K. Lilley, S. Hester, J. Howard, J. Rees, S. Russell, and D. St. Johnston. Genome-wide mapping and characterisation of protein expression and interaction in Drosophila melanogaster, using a hybrid piggyBac/P-element YFP gene trap system with tandem affinity tags. In *48th Annual Drosophila Research Conference*, pages 349+, Philadelphia, PA, March 2007.

[114] Stephan Saalfeld, Albert Cardona, Volker Hartenstein, and Pavel Tomančák. CATMAID: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics*, 25(15):1984–1986, August 2009.

[115] Stephan Saalfeld, Albert Cardona, Volker Hartenstein, and Pavel Tomančák. As-rigid-as-possible mosaicking and serial section registration of large ssTEM datasets. *Bioinformatics*, 26(12):i57–i63, June 2010.

[116] Dora D. Santagata, Tudor A. Fulga, Atanu Duttaroy, and Mel B. Feany. Oxidative stress mediates tau-induced neurodegeneration in Drosophila. *The Journal of Clinical Investigation*, 117(1):236–245, January 2007.

[117] Benjamin Schmid, Johannes Schindelin, Albert Cardona, Mark Longair, and Martin Heisenberg. A high-level 3D visualization API for Java and ImageJ. *BMC Bioinformatics*, 11(1):274+, 2010.

[118] A. Selverston. Modulation of circuits underlying rhythmic behaviors. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 176(2):139–147, February 1995.

[119] Kazunori Shinomiya, Keiji Matsuda, Takao Oishi, Hideo Otsuna, and Kei Ito. Flybrain neuron database: A comprehensive database system of the Drosophila brain neurons. *J. Comp. Neurol.*, 519(5):807–833, 2011.

[120] Xiaokun Shu, Varda Lev-Ram, Thomas J. Deerinck, Yingchuan Qi, Ericka B. Ramko, Michael W. Davidson, Yishi Jin, Mark H. Ellisman, and Roger Y. Tsien. A Genetically Encoded Tag for Correlated Light and Electron Microscopy of Intact Cells, Tissues, and Organisms. *PLoS Biol*, 9(4):e1001041+, April 2011.

[121] Stephen J. Smith. Circuit reconstruction tools today. *Current opinion in neurobiology*, 17(5):601–608, October 2007.

[122] Temple F. Smith and Michael S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, March 1981.

[123] C. Sommer, C. Straehle, U. Kothe, and F. A. Hamprecht. Ilastik: Interactive learning and segmentation toolkit. pages 230–233, March 2011.

[124] Hiroshi Suzuki, Tod R. Thiele, Serge Faumont, Marina Ezcurra, Shawn R. Lockery, and William R. Schafer. Functional asymmetry in Caenorhabditis elegans taste neurons and its computational role in chemotaxis. *Nature*, 454(7200):114–117, July 2008.

[125] M. Suzuki, T. Tsuji, and H. Ohtake. A model of motor control of the nematode with neuronal circuits. *Artificial Intelligence in Medicine*, 35(1-2):75–86, September 2005.

[126] Shin-Ya Y. Takemura, Zhiyuan Lu, and Ian A. Meinertzhagen. Synaptic circuits of the Drosophila optic lobe: the input terminals to the medulla. *The Journal of comparative neurology*, 509(5):493–513, August 2008.

[127] Stephen T. Thibault, Matthew A. Singer, Wesley Y. Miyazaki, Brett Milash, Nicholas A. Dompe, Carol M. Singh, Ross Buchholz, Madelyn Demsky, Robert Fawcett, Helen L. Francis-Lang, Lisa Ryner, Lai M. Cheung, Angela Chong, Cathy Erickson, William W. Fisher, Kimberly Greer, Stephanie R. Hartouni, Elizabeth Howie, Lakshmi Jakkula, Daniel Joo, Keith Killpack, Alex Laufer, Julie Mazzotta, Ronald D. Smith, Lynn M. Stevens, Christiana Stuber, Lory R. Tan, Richard Ventura, Alesa Woo, Irena Zakrajsek, Lora Zhao, Feng Chen, Candace Swimmer, Casey Kopczynski, Geoffrey Duyk, Margaret L. Winberg, and Jonathan Margolis. A complementary transposon tool kit for Drosophila melanogaster using P and piggyBac. *Nature Genetics*, 36(3):283–287, February 2004.

[128] Ulrich Thomas, Eunjoon Kim, Sven Kuhlendahl, Young H. Koh, Eckart D. Gundelfinger, Morgan Sheng, Craig C. Garner, and Vivian Budnik. Synaptic Clustering of the Cell Adhesion Molecule Fasciclin II by Discs-Large and its Role in the Regulation of Presynaptic Structure. *Neuron*, 19(4):787–799, October 1997.

[129] Ulrich Thomas, Bounpheng Phannavong, Bettina Müller, Craig C. Garner, and Eckart D. Gundelfinger. Functional expression of rat synapse-associated proteins SAP97 and SAP102 in Drosophila dlg-1 mutants: effects on tumor suppression and synaptic bouton structure. *Mechanisms of Development*, 62(2):161–174, March 1997.

[130] Srinivas C. Turaga, Kevin L. Briggman, Moritz Helmstaedter, Winfried Denk, and H. Sebastian Seung. Maximin affinity learning of image segmentation. *NIPS*, November 2009.

[131] Kiren Ubhi, Hassan Shaibah, Tracey Newman, David Shepherd, and Amritpal Mudher. A comparison of the neuronal dysfunction caused by Drosophila tau and human tau in a Drosophila model of tauopathies. *Invertebrate Neuroscience*, 7(3):165–171, September 2007.

[132] Peter Uetz and Robert E. Hughes. Systematic and large-scale two-hybrid screens. *Current Opinion in Microbiology*, 3(3):303–308, June 2000.

[133] Ranjith Unnikrishnan, Caroline Pantofaru, and Martial Hebert. Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, April 2007.

[134] Kannan U. Venkataraju, Antonio R. C. Paiva, Elizabeth Jurrus, and Tolga Tasdizen. Automatic markup of neural cell membranes using boosted decision stumps. In *ISBI'09: Proceedings of the Sixth IEEE international conference on Symposium on Biomedical Imaging*, pages 1039–1042, Piscataway, NJ, USA, 2009. IEEE Press.

[135] Leonie Welberg. Learning and memory: ORB2 marks the spot. *Nature Reviews Neuroscience*, 8(12):908–909, December 2007.

[136] John G. White, Eileen Southgate, J. Nichol Thomson, and Sydney Brenner. The structure of the ventral nerve cord of Caenorhabditis elegans. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 275(938):327–348, August 1976.

[137] Curtis W. Wittmann, Matthew F. Wszolek, Joshua M. Shulman, Paul M. Salvaterra, Jada Lewis, Mike Hutton, and Mel B. Feany. Tauopathy in Drosophila: Neurodegeneration Without Neurofibrillary Tangles. *Science*, 293(5530):711–714, July 2001.

[138] Shouzhen Xia, Tomoyuki Miyashita, Tsai-Feng Fu, Wei-Yong Lin, Chia-Lin Wu, Lori Pyzocha, Inn-Ray Lin, Minoru Saitoe, Tim Tully, and Ann-Shyn Chiang. NMDA Receptors Mediate Olfactory Learning and Memory in Drosophila. *Current Biology*, 15(7):603–615, April 2005.

[139] M. Yoshihara, A. W. Ensminger, and J. T. Littleton. Neurobiology and the Drosophila genome. *Functional & integrative genomics*, 1(4):235–240, March 2001.

[140] J. M. Young and J. D. Armstrong. Structure of the adult central complex in Drosophila: Organization of distinct neuronal subsets. *J. Comp. Neurol.*, 518(9):1500–1524, 2010.

[141] Jingkai Yu and Russell L. Finley. Combining multiple positive training sets to generate confidence scores for protein-protein interactions. *Bioinformatics*, 25(1):105–111, January 2009.

[142] Jingkai Yu, Svetlana Pacifico, Guozhen Liu, and Russell Finley. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9(1):461+, October 2008.

Appendix A

BrainTrap: a database of 3D protein expression patterns in the Drosophila brain

Seymour Knowles-Barley, Mark Longair, and Douglas Armstrong.
Published in *Database*, 2010 baq005. Accepted 25 February 2010.

Appendix B

Biologically inspired EM image alignment and neural reconstruction

Seymour Knowles-Barley, Nancy J Butcher, Ian A Meinertzhagen, and Douglas Armstrong.

Published in *Bioinformatics*. First published online: July 8, 2011.